Integrating Video and Text: A Balanced Approach to Multimodal Summary Generation and Evaluation

Galann Pennec[∞,◊,♡] Zhengyuan Liu[◊]
Nicholas Asher[§,♡] Philippe Muller[∞,♡] Nancy F. Chen[◊]
[∞]IRIT, University of Toulouse, France [◊]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[♡]CNRS@CREATE, Singapore [§]CNRS, IRIT, France
galann.pennec@cnrsatcreate.sg, {liu_zhengyuan,nfychen}@i2r.a-star.edu.sg
{nicholas.asher,philippe.muller}@irit.fr

Abstract

Vision-Language Models (VLMs) often struggle to balance visual and textual information when summarizing complex multimodal inputs, such as entire TV show episodes. In this paper, we propose a zero-shot video-to-text summarization approach that builds its own screenplay representation of an episode, effectively integrating key video moments, dialogue, and character information into a unified document. Unlike previous approaches, we simultaneously generate screenplays and name the characters in zero-shot, using only the audio, video, and transcripts as input. Additionally, we highlight that existing summarization metrics can fail to assess the multimodal content in summaries. To address this, we introduce MFACTSUM, a multimodal metric that evaluates summaries with respect to both vision and text modalities. Using MFACTSUM, we evaluate our screenplay summaries on the SummScreen3D dataset, demonstrating superiority against state-of-the-art VLMs such as Gemini 1.5 by generating summaries containing 20% more relevant visual information while requiring 75% less of the video as input.

1 Introduction

VLMs (Dubey et al., 2024; Cheng et al., 2024; OpenAI, 2024) still struggle to effectively balance both vision and text modalities in their answers, sometimes neglecting or completely ignoring one input modality over the other (Zhang et al., 2024d; Nishimura et al., 2024; Shen et al., 2024; Park et al., 2024). This challenge is particularly present in video-text tasks, where datasets often suffer from poor annotations and limited diversity, restricting the model's ability to bridge the gap between vision and text (Hua et al., 2024). Furthermore, there are significant uncertainties regarding how VLMs handle long multimodal contexts, such as hour-long videos or sequences of hundreds of images (Fu et al., 2024; Song et al., 2024; Wang et al., 2024c; Zhou et al., 2024; Zhang et al., 2024c).

To address these challenges, storytelling approaches have emerged as promising alternatives. These methods aim to improve multimodal understanding by first generating textual descriptions from videos (e.g., scripts, screenplays) before applying them to various downstream tasks (Wu et al., 2024; Zhang et al., 2024a; Bhattacharyya et al., 2023). However, little is known about how storytelling approaches compare to state-of-the-art VLMs in summarizing long videos and transcripts.

In this paper, we propose a zero-shot multimodal pipeline (Figure 1) to summarize long videos, such as TV show episodes, by building our own textual screenplays. We argue that screenplays provide a natural way for both human users (e.g. actors) and Large Language Models (LLMs) to access and interpret all the multimodal content of an episode into a unified document, including dialogue, character interactions, emotions, behaviors and scene locations. Those screenplay representations are reusable across different tasks, thereby limiting the need to reencode the whole video every time.

Unlike previous storytelling approaches, we both generate video captions and reidentify characters at the same time by prompting a VLM in zero-shot on the audio, video and transcripts. Additionally, we reduce the generation cost of our screenplays while preserving their multimodal richness. We do this by identifying and summarizing key moments from the video (Figure 1), since we observe that such moments often align with natural pauses in the dialogue.

Moreover, we find that existing summarization metrics tend to overlook visual content understanding, as such information is often less present in TV show summaries. This motivates us to introduce a new evaluation metric, MFACTSUM, designed to assess the multimodal fidelity of video-to-text summaries. Such a metric is crucial for the task, as

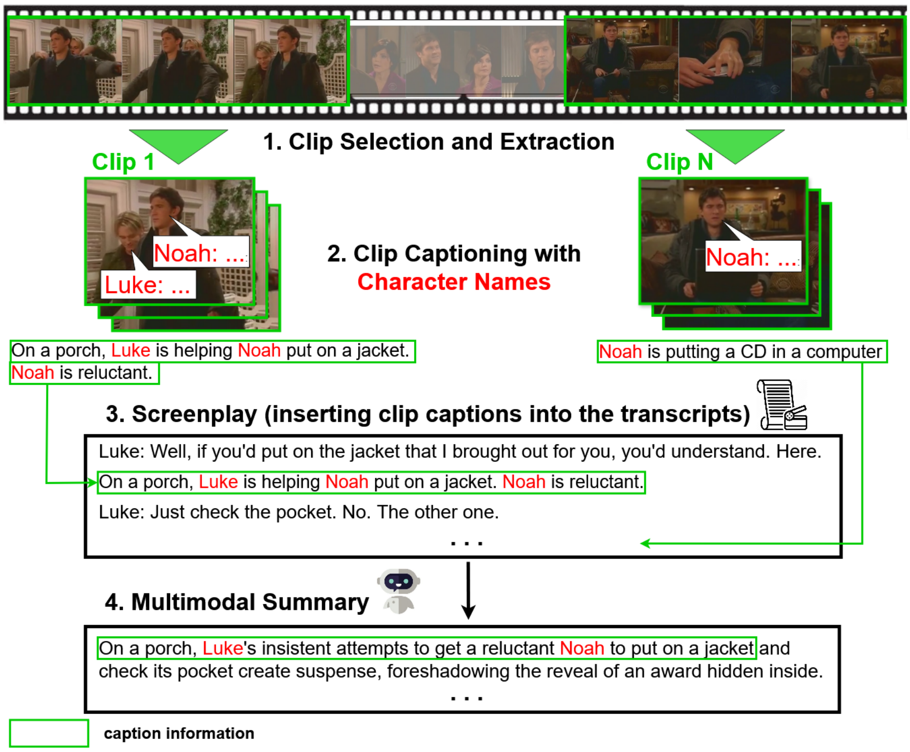
1. Clip Selection and Extraction
Clip 1
Clip N
Noah: ...
Luke: ...
Noah: ...
2. Clip Captioning with Character Names
On a porch, Luke is helping Noah put on a jacket.
Noah is reluctant.
Noah is putting a CD in a computer
3. Screenplay (inserting clip captions into the transcripts)
Luke: Well, if you'd put on the jacket that I brought out for you, you'd understand. Here.
On a porch, Luke is helping Noah put on a jacket. Noah is reluctant.
Luke: Just check the pocket. No. The other one.
...
4. Multimodal Summary
On a porch, Luke's insistent attempts to get a reluctant Noah to put on a jacket and check its pocket create suspense, foreshadowing the reveal of an award hidden inside.
...
□ caption information

Figure 1: **Our zero-shot pipeline for summarizing long TV show episodes** 1) We select and extract important video clips from the whole episode 2) We provide a video clip and its corresponding transcripts to a VLM for caption generation while reidentifying featured characters 3) Screenplays are built by inserting the clip captions at the correct timestamp in the full episode transcripts 4) We finally perform summarization of those screenplays

text modality alone cannot fully capture emotions, actions, and locations that provide essential context to an episode. Inspired by factual consistency metrics such as FactScore (Min et al., 2023) and PRISMA (Mahon and Lapata, 2024), our metric evaluates the recall of both visual facts (video related) and textual facts (transcripts related) within a generated summary. By separately assessing visual and textual understanding, MFACTSUM provides a balanced evaluation of a summarization system's multimodal capabilities.

In summary, our contributions are as follows:

- We introduce a zero-shot multimodal approach (Figure 1) that summarizes TV shows by building a screenplay document integrating key video moments, transcripts, and character information.
- We propose a new multimodal metric, MFACTSUM to evaluate how well a summarization system equally captures both textual and visual information.
- We evaluate our pipeline on SummScreen3D[1] (Papalampidi and Lapata, 2023) containing long soap opera episodes. Our results show that our screenplay-based summaries retrieve 20% more visual information than Gemini 1.5 Pro while using about 75% less of the video as input.

Through this work, we aim to improve summarization models and evaluation protocols by making them more balanced across modalities, ultimately advancing the field of multimodal summarization.

2 Related Work

Multimodal Summarization of Movies and TV Shows The task can be performed in two different ways depending on the nature of the summary. While the output summary can be a video (a recap or trailer) (Singh et al., 2024; Papalampidi et al., 2021; Chen et al., 2024a), our study instead focuses on generating textual summaries from a TV show episode.

Originally, textual summaries were produced given scripts or screenplays of the movie or TV

[1] https://github.com/ppapalampidi/long_video_summarization

show episode (Saxena and Keller, 2024; Gorinski and Lapata, 2015). More recently, multimodal summarization datasets like SummScreen3D (Papalampidi and Lapata, 2023), presented in detail in Section 5.1, have led the way to new Vision-Language approaches for the task (Papalampidi and Lapata, 2023; Mahon and Lapata, 2024).

Yet, little is known about how these models leverage both modalities when summarizing hour-long TV show episodes. We therefore propose MFACTSUM as a new multimodal evaluation strategy.

Storytelling Methods Movie audio description (Rohrbach et al., 2015; Soldan et al., 2022) or movie screenplays (Gorinski and Lapata, 2015; Saxena and Keller, 2024) fall into the same category of documents that relate the visual elements of a movie and put them back in the context of the story for a complete and accurate multimodal understanding.

When available for a movie, screenplays or other forms of narration can serve as a basis to perform any downstream task such as movie summarization (Saxena and Keller, 2024; Gorinski and Lapata, 2015; Sang et al., 2022; Gorinski and Lapata, 2018; Reboud et al., 2023; Huang et al., 2020).

Some recent approaches known as storytelling methods even automatically generate their own screenplays or textual descriptions of a movie to later perform video understanding (Wu et al., 2024; Zhang et al., 2024a; Bhattacharyya et al., 2023), trailer prediction (Chen et al., 2024a) or even allow the user to have a conversation over an entire movie (Lin et al., 2023). In this paper, we generate our own screenplays from the input video and transcripts and use them for summarizing entire TV show episodes (see Section 3).

Character Identification Characters constitute a central part of any story. Character identification in a video helps the models better connect what is displayed on screen to the actual conversations, eventually improving multimodal understanding of movies and TV shows.

Applications of character identification in story understanding include Video Question Answering (VideoQA) (Geng et al., 2020; Lei et al., 2020; Choi et al., 2021), movie audio description (Pini et al., 2019; Han et al., 2023, 2024; Wang et al., 2024a; Xie et al., 2024; Raajesh et al., 2024) and movie summarization (Sang and Xu, 2010; Tran et al., 2017).

The task can be performed based on the video coupled with metadata (e.g. from IMDb[2]) (Han et al., 2023, 2024; Xie et al., 2024) but can also be treated thanks to existing annotations such as speaker names from movie transcripts (Geng et al., 2020).

Unlike previous approaches, we perform character identification and video captioning in one go without relying on face annotations from IMDb or training any new algorithms, by prompting a VLM on the audio, video and transcripts (see Section 3.2).

Summary Evaluation Common strategies for evaluating artificial summaries include word n-gram comparisons such as ROUGE (Lin, 2004) or METEOR (Banerjee and Lavie, 2005), neural-based evaluation like BertScore (Zhang et al., 2020), factual consistency evaluation (Laban et al., 2022; Kryscinski et al., 2020; Maynez et al., 2020; Krishna et al., 2023) or QA-based evaluation (Durmus et al., 2020; Fabbri et al., 2022).

Most of the above metrics have been widely used in multimodal summary evaluation (Papalampidi and Lapata, 2023; Mahon and Lapata, 2024). Yet, they have not been designed to assess the multimodal richness of summaries. We therefore propose a metric, MFACTSUM, specifically designed to evaluate the multimodal content in video and text summaries (see Section 4).

While there has been some effort in balancing and/or aligning modalities in video summarization models (Shen et al., 2024; Hua et al., 2024; Liu et al., 2020; Lin et al., 2024) and even making their output interpretable with respect to every modality (Tian et al., 2018), there has been, to our knowledge, no specific work on developing balanced summarization metrics for the multimodal setting.

In addition, while some progress has been made in developing metrics for image and text summarization (Jing et al., 2024; Wan and Bansal, 2022; Hessel et al., 2021; Zhang et al., 2023; Zhu et al., 2018, 2020), no such metrics have been proposed yet for videos to the best of our knowledge.

3 Screenplay Generation and Summarization

We describe our zero-shot multimodal summarization pipeline in Fig.1, which takes the whole video and transcript as input. The first step is to select

[2] `https://www.imdb.com`

and extract all the clips of interest from the whole video (Section 3.1). We then generate captions for each clip using a VLM while identifying the main characters appearing in them (Section 3.2). We build the screenplay by aligning the resulting clip captions in time together with the transcripts. We finally feed the screenplays to an LLM for summarization (Section 3.3).

3.1 Clip Selection

Selecting important clips for long video summarization is a challenging task. When it comes to VideoQA, some approaches (Yu et al., 2023) train a model to predict the most important keyframes given the question to answer.

As we want to keep our pipeline training-free, we select all video clips that occur during a pause in the dialogue when no speech is detected. We ensure that all such video clips are extended to a minimum duration of 10 seconds to give enough context for further captioning.

Our clip selection strategy has two motivations:

- Silent scenes from a video often highlight key visual moments and actions impacting the episode storyline. For instance, in *As the World Turns*, a deeply moving moment occurs when Noah inserts the CD of his own movie into a computer (Figure 1).
- In Audio Description, the narrator's interventions are usually placed during such breaks in the dialogue suggesting the importance these moments have in the unfolding of an episode.

We further validate our proposed criterion in Section 5.5. With this criterion, we reduce the cost of our experiments on SummScreen3D (Section 5) by using about 25% of the entire video as input.

3.2 Clip Captioning with Character Names

We feed all video clips to a VLM for captioning. We also provide the corresponding transcripts from SummScreen3D so that the model can reidentify the characters present in the video clips and explicitly state their names in the produced captions. This is possible because each transcript line contains the name of the speaker. The model can therefore easily deduce which characters appear at a given time in the clips by matching each transcript line with the corresponding audio and video. We further validate the performance of our character identification in Section 5.5. Detailed captioning prompts are provided in Appendix A.1.

3.3 Screenplay Summarization

The screenplay is built by interleaving the clip captions generated in Section 3.2 and the transcript utterances together in time. To enhance clarity for the LLM, we prepend `'Caption:'` to every clip caption within the screenplay. We end up with a document where every clip caption has been inserted at the correct timestamp in the transcripts. We further summarize those screenplays with a LLM, using a prompt specifically tailored for the multimodal task, asking the model to pay attention to important visual cues like characters' actions or scene locations (see Appendix A.2.1).

4 Multimodal Summary Evaluation with MFACTSUM

When it comes to movie and TV show summarization, the reference summaries are often imbalanced (see Appendix C) as they usually contain more information regarding what can be heard or read from the transcripts (text modality) than what can be seen (vision modality). Due to this imbalance, most of the evaluation metrics in section 2, like ROUGE and factual consistency metrics, do not account well for multimodal understanding.

Our proposed metric, MFACTSUM, assesses the effectiveness of a multimodal summarization system in balancing and integrating information from both the video and transcripts. MFACTSUM therefore gives a new insight on multimodal summary evaluation that cannot provide most traditional summarization metrics. Such a metric is essential to properly assess the inclusion in the summary of all the important video information absent from the transcripts such as performed actions (e.g. `'Brooke kisses Ridge'`), emotions (e.g. `'Beth is hysterical'`) or settings (e.g. `'sitting in a wheelchair'`).

4.1 Metric Design

Our metric is a factual consistency metric similar to Factscore (Min et al., 2023) and PRISMA (Mahon and Lapata, 2024). PRISMA decomposes the groundtruth summary into its own facts, each fact conveying a single piece of information (roughly equivalent to a simple clause, see examples in Appendix A.3.1). Completeness of a summary is then evaluated with a "fact recall", i.e. the ratio of facts from the groundtruth summary being supported by the predicted summary.

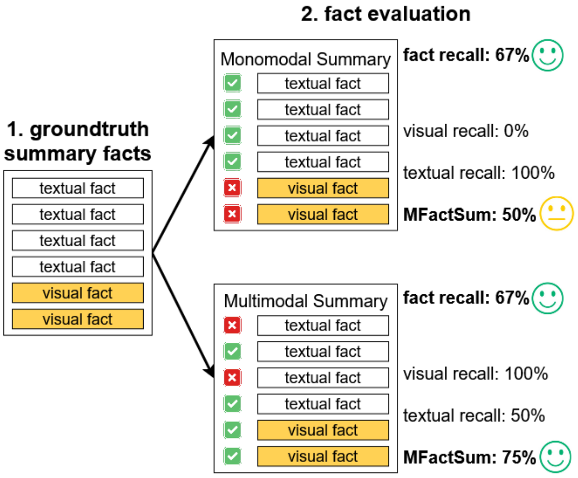

Figure 2: **For the same number of recalled facts, MFACTSUM favors the multimodal summary**. 1) We identify all the visual and textual facts within the groundtruth summary. To be consistent with our study dataset, we choose a groundtruth summary with an uneven distribution of textual and visual facts. 2) Although PRISMA's fact recall score is the same for both the multimodal and monomodal summaries, our metric MFACTSUM favors the multimodal summary over the monomodal one.

As opposed to PRISMA, our metric is not biased towards a single modality as it gives equal importance to visual facts (facts referring to the video) and textual facts (facts referring to the transcripts) in the recall computation. We provide an example comparison of the two metrics in Figure 2. MFACTSUM ultimately provides deeper insights into how well an approach understands a video from the perspective of what can be heard (text modality) and what can be seen (vision modality).

Unlike PRISMA, which also considers fact precision for the relevance of a summary, our metric MFACTSUM is recall-based. This is because our study focuses on evaluating the unimodal bias in summarization systems. In other words, we assess whether or not a summary recalls relevant information from both modalities.

4.2 Metric Computation

Our metric computation involves three different steps: **Fact Identification**, **Visual Fact Classification** and **Fact Evaluation**. The only step that is kept unchanged compared to PRISMA design is the identification of facts.

Fact Identification We split the groundtruth summary into its own sentences. We further identify the facts within each sentence by prompting an LLM in a few-shot setting. Our few-shot examples can be found in Appendix A.3.1. At this stage, we end up with a list of groundtruth summary facts on which we can later perform our evaluations.

Visual Fact Classification We classify whether a fact from the groundtruth is `Visual` (related to the video) or `Textual` (related to the dialogue between the characters). We propose two simple steps to separate `Visual` from `Textual` facts.

1. A fact is classified as `Visual` if it cannot be inferred from the transcripts alone. We therefore prompt the LLM in zero-shot to answer whether the fact is supported by the transcripts. Our prompt is given in Appendix A.3.2.

2. We manually hand-label a list of facts in A.3.2 as `Visual` or `Textual` and use them as few-shot examples for the classification task.

Each of the above steps is performed independently, and both must classify a fact as `Visual` for it to be considered as such. In all other cases, the fact is classified as `Textual`. We further validate our classification method in Section 5.5.

Fact Evaluation In PRISMA, the fact recall (`fact-rec`) is defined by the ratio of facts from the groundtruth summary that are supported by the predicted summary.

Instead, we compute the recall for both visual and textual facts separately. This allows us to define MFACTSUM as the average of visual recall (`vis-rec`) and textual recall (`text-rec`).

$$\text{MFACTSUM} = \frac{\texttt{vis-rec} + \texttt{text-rec}}{2}.$$

Doing the mean at this stage sets the same importance to vision and text modalities in the metric computation.

5 Experiments

5.1 Datasets

SummScreen3D is a video-to-text summarization dataset (Papalampidi and Lapata, 2023) of 5421 episodes of about 30 to 60 minutes each from famous soap operas (As the World Turns, The Bold and the Beautiful, . . .). It includes rather long transcripts (about 6K tokens on average), videos and multiple summaries for each episode. The validation and test splits contain 296 episodes each.

SummScreen3D is a simple extension of SummScreen (Chen et al., 2022) into a multimodal summarization dataset by adding full-length videos to the already existing transcripts and reference summaries. The summaries in SummScreen3D are also highly multimodal as they contain information referring to both the episode video and transcripts.

Aligning in time every line from the transcripts to its corresponding frames from the video is crucial for an accurate multimodal understanding of a TV show episode and for generating our screenplays. We rely on previous work (Mahon and Lapata, 2024) to achieve this.

5.2 Evaluation Metrics

We report in Table 1 the multimodal performance based on our new metrics (`vis-rec`, `text-rec` and MFACTSUM) presented in Section 4. For comparison, we report the simple fact recall score (`fact-rec`) as defined in PRISMA in which no reweighting of visual and textual facts is performed. We use the lighter model Gemini 1.5 Flash as the base model for all our fact-based evaluations.

We also include the average ROUGE-1 (r1), ROUGE-2 (r2) and ROUGE-Lsum (rlsum) as given by the python-rouge package and include two additional metrics, METEOR and BertScore, in Appendix D for further comparison.

As multiple reference summaries are provided for each episode in SummScreen3D, we take the maximum ROUGE score against all references. When doing our fact-based evaluation, we only use the groundtruth summaries from `soap_central` found in SummScreen3D, as they are longer on average and thus more likely to contain visual facts. More precisely, we identify 20 visual facts on average in those summaries making them about 14.5% of the total number of facts. We filter out the 28 episodes from the test set for which no `soap_central` summary is provided.

We always report the average word count (`avg-len`) of a system's summaries in Table 1, as we are aware that both visual and textual recall can increase with summary length. When comparing any two summarization systems in Section 5.6, we therefore always make sure that their summary lengths are comparable.

5.3 Implementation Details

We generate screenplays for all the 296 episodes from SummScreen3D test split using the pipeline described in Section 3. We use either Gemini 1.5 Pro (Reid et al., 2024) or Qwen2-VL-72B (Wang et al., 2024b) as both the captioning and screenplay summarization models. We pick up those two models as they are currently among the best performing VLMs according to multiple long-form video benchmarks (Fu et al., 2024; Mangalam et al., 2023; Wang et al., 2024c) and also for their long context abilities. We also run various baselines (see Section 5.4) on the same split for comparison. Due to the high API costs, we only provide results from a single run.

5.4 Comparison Baselines

We compare our approach to existing works from the literature including finetuned models on SummScreen3D and zero-shot pipelines for long video understanding. As the video-to-text summarization of movies and TV shows is a recent task, very few finetuned end-to-end approaches have been proposed. In addition, we do not compare to existing works on Audio Description (Xie et al., 2024; Chu et al., 2024; Zhang et al., 2024b) because they operate on specific selected video segments rather than the whole video as input. For that reason, we also include state-of-the-art open and private VLMs such as Gemini 1.5 Pro and Qwen2-VL-72B as comparison baselines in order to perform the summarization task end-to-end. We also tried other VLMs (Liu et al., 2024; Cheng et al., 2024; Chen et al., 2024b) but did not include them in our experiments as we found them unsuitable for end-to-end analysis of hour-long videos and transcripts.

Modular-Kosmos (Mahon and Lapata, 2024) A multimodal approach finetuned on SummScreen3D. It decomposes an episode into its own scenes to further perform summarization and generate video captions for each of them. A higher-level BART (Lewis et al., 2020) is then finetuned to fuse both scene summaries and video captions into one global summary for the entire episode.

VLog An open-source tool[3] that converts a video into a long document thanks to vision captioners and a speech recognition model. The document is later fed to ChatGPT (OpenAI, 2023) to start a conversation. To further generate the summaries with VLog, we use the prompt in Appendix A.2.2 that is very similar to the one used for summarizing our own screenplays.

Gemini 1.5 Pro and Qwen2-VL-72B (video) We prompt both models on full videos and transcripts in an end-to-end fashion for multimodal summarization. We extract the maximum number of frames from the videos that are allowed by each

[3] https://github.com/showlab/VLog

	vis-rec	text-rec	MFS	fact-rec	r1	r2	rlsum	avg-len
multimodal baselines								
Modular-Kosmos (Mahon and Lapata, 2024)	7.39	19.56	13.48	17.90	44.86	**11.83**	42.97	314.0
VLog	7.77	15.66	11.72	14.62	25.99	3.11	24.69	314.0
Qwen2-VL-72B (no video)	16.16	38.0	27.08	35.0	40.6	8.97	39.28	718.5
Qwen2-VL-72B (video)*	23.69	37.50	30.60	35.61	38.06	8.11	36.95	889.0
Screenplay Summary (Qwen2-VL-72B)	24.43	35.45	29,94	33.93	36.50	7.23	35.51	749.5
Gemini 1.5 Pro (no video)	21.54	42.44	31.99	39.52	41.52	9.04	40.06	573.9
Gemini 1.5 Pro (video)*	27.48	43.00	35.24	40.87	**46.67**	11.77	**44.99**	688.3
Screenplay Summary (Gemini 1.5 Pro)	**33.04**	**45.12**	**39.08**	**43.53**	40.23	8.57	38.82	601.1

Table 1: **Evaluation results on SummScreen3D.** We report the visual recall (`vis-rec`), textual recall (`text-rec`) and MFactSum denoted as MFS. For comparison, we also include ROUGE-1 (r1), ROUGE-2 (r2), ROUGE-Lsum (rlsum) and the simple fact recall (`fact-rec`). The average summary word count is denoted by (`avg-len`). Best results are in **bold**. * indicates the VLM is prompted on the full video and transcripts in an end-to-end fashion using the maximum number of frames allowed by the API.

model's API which is 1 frame per second for Gemini[4] and 250 frames for Qwen2-VL[5]. For fairness, we use a prompt (see Appendix A.2.3) that is very similar to the one we used for our own screenplay summarization approach as we assume it leads to richer multimodal summaries.

Gemini 1.5 Pro and Qwen2-VL-72B (no video) We generate summaries from the transcripts alone using each model's API. We provide the prompt for this setup in Appendix A.2.4.

5.5 Human Evaluation

We conduct human validation of various components of both our summarization pipeline and evaluation metric (see Appendix B).

MFactSum Our metric robustly identifies visual facts from groundtruth summaries, achieving an accuracy of about 86.3% on 573 facts (Appendix B.1).

Clip Selection Our selected video clips (Section 3.1) capture approximately two thirds of the visual information present in groundtruth summaries, which is about 1.9 times more than what is achieved using randomly selected clips (Appendix B.2).

Character Identification We found a 77.4% overlap with the human annotator in identifying characters across 174 video clips (Appendix B.3). This result is on par with prior work.

5.6 Results

Our screenplay summaries are more multimodal than VLM summaries. Our screenplay summaries produced with Gemini 1.5 Pro recall about 20% more visual facts than Gemini 1.5 Pro prompted end-to-end on the full videos and transcripts, while still maintaining a comparable ability to recall textual information (Table 1). We observe a similar trend, to a lesser extent, when substituting Gemini 1.5 Pro for Qwen2-VL-72B.

This suggests that, for the task, relying on our screenplays as a multimodal representation, allows us to outperform a VLM model prompted end-to-end. Note that this improvement in terms of visual recall is not due to a difference in summary length (`avg-len`) as our screenplay summaries are actually shorter.

In Appendix E, we compare the amount of visual content retrieved by the different tested models in their summary for a single episode.

Traditional metrics are not always enough for multimodal summary evaluation. Traditional summarization metrics like ROUGE in Table 1, even METEOR and BertScore in Appendix D, are often surprisingly higher for summaries that are less multimodal. In particular,

- Our screenplay summaries are 20% more visual than Gemini 1.5 Pro prompted end-to-end on the full videos, while they are still about 6 ROUGE points behind.

- A finetuned model like Modular-Kosmos has better ROUGE scores than most other approaches. Yet, its visual recall is the lowest of all compared to its textual recall, suggesting that its multimodal ability is weak.

This phenomenon can be explained by the fact that a multimodal summary is inherently more abstractive than a monomodal one, which ultimately

[4] `https://aistudio.google.com/`

[5] `https://www.alibabacloud.com/en/product/modelstudio`

	vis-rec	text-rec	MFS	fact-rec	r1	r2	rlsum	avg-len
w/o handcrafted prompt	20.55	40.66	30.61	37.96	**40.91**	**8.78**	**39.54**	609.9
w/o character ident.	**34.35**	43.32	38.84	42.14	39.97	8.26	38.58	576.5
Screenplay Summary (Gemini 1.5 Pro)	33.04	**45.12**	**39.08**	**43.53**	40.22	8.65	38.90	601.1

Table 2: **Ablation results for our screenplay summarization pipeline using Gemini 1.5 Pro as the based model.** We report the visual recall (`vis-rec`), textual recall (`text-rec`) and MFactSum denoted as MFS. For comparison, we also include ROUGE-1 (r1), ROUGE-2 (r2), ROUGE-Lsum (rlsum) and the simple fact recall (`fact-rec`). The average summary word count is denoted by (`avg-len`). Best results are in **bold**.

negatively affects all the metrics that rely on exact lexical matching such as ROUGE or METEOR.

MFactSum better reflects multimodality in summaries. We compare across all metrics (Table 1) the scores between a multimodal summary like ours and a monomodal summary (no video as input). The conclusions here are the same for both Gemini 1.5 Pro and Qwen2-VL.

When no video is used as input, both MFactSum and the visual recall experience the largest drop. On the other hand, metrics such as the simple fact recall (`fact-rec`) either drop a little or even increase like ROUGE. This means that MFactSum better assesses the multimodal ability of a summarization system.

Note that the visual recall found for the monomodal summary is however greater than 0 (Table 1). We attribute this to potential inaccuracies in fact evaluation, as some facts may be ambiguous in relation to the overall context of an episode.

5.7 Ablations

We perform various ablations and draw the following conclusions from our results in Table 2.

Instruction Tuned LLMs such as Gemini do not naturally output multimodal summaries. Simply asking the LLM to summarize the screenplay is not enough to capture its multimodal content. To demonstrate this point we replace our own handcrafted prompt from Section 3.3 with a much simpler prompt that only asks to summarize (see Appendix A.2.4). We discover that, by doing so, the visual recall drops by about 40% (Table 2). Note that this drop is not caused by a difference in summary length (`avg-len`) between the two compared systems. This suggests that Instruction Tuned models such as Gemini 1.5 Pro are naturally biased into generating monomodal summaries and need to be specifically prompted to retrieve important visual information from the screenplays. Note that this bias can also be found in a finetuned model such as Modular-Kosmos (see Appendix C).

Character identification has little impact on our pipeline performance. Removing the character identification part from our pipeline has little impact on both ROUGE and factual consistency metrics (Table 2). One possible reason is that character identities can often be inferred from the screenplay itself, thanks to the dialogue context surrounding each clip caption. Since our screenplays aim to provide a comprehensive representation of an episode, we decide to still include the character identification component within our main pipeline. The prompts used for this ablation are in Appendix A.1.

6 Conclusion

In this paper, we introduce a zero-shot video-to-text summarization approach that constructs a multimodal screenplay of an episode, effectively integrating important visual cues, transcripts, and character information. Unlike previous approaches, we recognize characters in the video while producing video captions simultaneously using only the audio, video, and transcripts as input, avoiding the need for extra face annotations (IMDb).

We also propose a new multimodal metric, MFactSum, which better reflects the multimodal fidelity of summaries than traditional metrics according to our experimental results. Evaluation with MFactSum shows that LLMs and finetuned models are naturally biased and tend to include less visual content in their summaries. In contrast, using our screenplay representation leads to summaries that not only incorporate 20% more relevant visual information than state-of-the-art vision-language models like Gemini 1.5 Pro but also require 75% less of the video as input.

Future works could include expanding our evaluation metrics to other multimodal generative tasks and domains, and this can also contribute to explore methods to reduce modality biases through better representation learning and alignment.

Limitations

The motivation behind our proposed clip selection strategy comes from the audio description of movies or TV shows 3.1. Different strategies should be explored to identify key moments for other video domains while keeping the overall inference cost low.

Despite the effectiveness of our evaluation approach, we identify two main challenges that still need to be addressed.

- **Fact Evaluation:** LLMs might still generate some errors in the evaluation as the facts can sometimes be ambiguous with respect to the whole context of the episode.
- **Visual Fact Classification:** Leveraging the episode video for identifying visual facts is currently out of reach as current VLMs still produce inaccuracies when prompted on long videos (e.g. long video question answering).

Finally, the financial cost per episode of generating our screenplay summaries and evaluating them using MFactSum is approximately 1.25 $ per episode. We also estimate the total cost of producing all the results presented in this paper to be approximately 1000 $. In particular, our fact-based evaluation with MFactSum involves multiple LLM queries with repeated long contexts, further increasing inference costs.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Aanisha Bhattacharyya, Yaman Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. 2023. A video is worth 4096 tokens: Verbalize story videos to understand them in zero shot. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9822–9839. Association for Computational Linguistics.

Brian Y. Chen, Xiangyuan Zhao, and Yingnan Zhu. 2024a. Personalized video summarization by multimodal video understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 4382–4389. ACM.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. Summscreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8602–8615. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *CoRR*, abs/2404.16821.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Min Su Lee, and Byoung-Tak Zhang. 2021. Dramaqa: Character-centered video story understanding with hierarchical QA. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 1166–1174. AAAI Press.

Peng Chu, Jiang Wang, and Andre Abrantes. 2024. LLM-AD: large language model based audio description system. *CoRR*, abs/2405.00983.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Esin Durmus, He He, and Mona T. Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5055–5070. Association for Computational Linguistics.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2587–2601. Association for Computational Linguistics.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075.

Shijie Geng, Ji Zhang, Zuohui Fu, Peng Gao, Hang Zhang, and Gerard de Melo. 2020. Character matters: Video story understanding with character-aware relations. *CoRR*, abs/2005.08646.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1066–1076. The Association for Computational Linguistics.

Philip John Gorinski and Mirella Lapata. 2018. What's this movie about? A joint neural network architecture for movie content analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1770–1781. Association for Computational Linguistics.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. Autoad II: the sequel - who, when, and what in movie audio description. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 13599–13609. IEEE.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. Autoad III: the prequel - back to the pixels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18164–18174. IEEE.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2024. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *CoRR*, abs/2404.12353.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 709–727. Springer.

Liqiang Jing, Jingxuan Zuo, and Yue Zhang. 2024. Fine-grained and explainable factuality evaluation for multimodal summarization. *CoRR*, abs/2402.11414.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1642–1661. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Trans. Assoc. Comput. Linguistics*, 10:163–177.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020. TVQA+: spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8211–8225. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2024. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Trans. Multim.*, 26:5548–5560.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, Ce Liu, and Lijuan Wang. 2023. MM-VID: advancing video understanding with gpt-4v(ision). *CoRR*, abs/2310.19773.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1834–1845. Association for Computational Linguistics.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. 2024. NVILA: efficient frontier visual language models. *CoRR*, abs/2412.04468.

Louis Mahon and Mirella Lapata. 2024. A modular approach for multimodal summarization of TV shows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8272–8291. Association for Computational Linguistics.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.

Taichi Nishimura, Shota Nakada, and Masayoshi Kondo. 2024. On the audio hallucinations in large audio-video language models. *CoRR*, abs/2401.09774.

OpenAI. 2023. Chatgpt-3.5. Accessed: 2025-02-13.

OpenAI. 2024. Hello gpt-4o. Accessed: 2024-11-6.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2021. Movie summarization via sparse graph construction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13631–13639. AAAI Press.

Pinelopi Papalampidi and Mirella Lapata. 2023. Hierarchical3d adapters for long video-to-text summarization. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1267–1290. Association for Computational Linguistics.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. 2024. Assessing modality bias in video question answering benchmarks with multimodal large language models. *CoRR*, abs/2408.12763.

Stefano Pini, Marcella Cornia, Federico Bolelli, Lorenzo Baraldi, and Rita Cucchiara. 2019. M-VAD names: a dataset for video captioning with naming. *Multim. Tools Appl.*, 78(10):14007–14027.

Haran Raajesh, Naveen Reddy Desanur, Zeeshan Khan, and Makarand Tapaswi. 2024. Micap: A unified model for identity-aware movie descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14011–14021. IEEE.

Alison Reboud, Ismail Harrando, Pasquale Lisena, and Raphaël Troncy. 2023. Stories of love and violence: zero-shot interesting events' classification for unsupervised TV series summarization. *Multim. Syst.*, 29(6):3951–3969.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3202–3212. IEEE Computer Society.

Jitao Sang and Changsheng Xu. 2010. Character-based movie summarization. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 855–858. ACM.

Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey M. Stanton. 2022. MBTI personality prediction for fictional characters using movie scripts. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6715–6724. Association for Computational Linguistics.

Rohit Saxena and Frank Keller. 2024. Moviesum: An abstractive summarization dataset for movie screenplays. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 4043–4050. Association for Computational Linguistics.

Yuhan Shen, Linjie Yang, Longyin Wen, Haichao Yu, Ehsan Elhamifar, and Heng Wang. 2024. Exploring the role of audio in video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 2090–2100. IEEE.

Aditya Kumar Singh, Dhruv Srivastava, and Makarand Tapaswi. 2024. "previously on..." from recaps to story summarization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13635–13646. IEEE.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. MAD: A scalable dataset for language grounding in videos from movie audio descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5016–5025. IEEE.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. Milebench: Benchmarking mllms in long context. *CoRR*, abs/2404.18532.

Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. 2018. An attempt towards interpretable audio-visual video captioning. *CoRR*, abs/1812.02872.

Quang Dieu Tran, Dosam Hwang, O-Joun Lee, and Jai E. Jung. 2017. Exploiting character networks for movie summarization. *Multim. Tools Appl.*, 76(8):10357–10369.

David Wan and Mohit Bansal. 2022. Evaluating and improving factuality in multimodal abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9632–9648. Association for Computational Linguistics.

Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. 2024a. Contextual AD narration with interleaved multimodal sequence. *CoRR*, abs/2403.12922.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024c. Lvbench: An extreme long video understanding benchmark. *CoRR*, abs/2406.08035.

Yongliang Wu, Bozheng Li, Jiawang Cao, Wenbo Zhu, Yi Lu, Weiheng Chi, Chuyun Xie, Haolin Zheng, Ziyue Su, Jay Wu, and Xu Yang. 2024. Zero-shot long-form video understanding through screenplay. *CoRR*, abs/2406.17309.

Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. Autoad-zero: A training-free framework for zero-shot audio description. *CoRR*, abs/2407.15850.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. In

Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024a. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 21715–21737. Association for Computational Linguistics.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2024b. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13647–13657. IEEE.

Litian Zhang, Xiaoming Zhang, Ziming Guo, and Zhipeng Liu. 2023. Cisum: Learning cross-modality interaction to enhance multimodal semantic coverage for multimodal summarization. In *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, pages 370–378. SIAM.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024c. Long context transfer from language to vision. *CoRR*, abs/2406.16852.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yifan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024d. Debiasing multimodal large language models. *CoRR*, abs/2403.05262.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. MLVU: A comprehensive benchmark for multi-task long video understanding. *CoRR*, abs/2406.04264.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4154–4164. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9749–9756. AAAI Press.

A Prompts

A.1 Clip captioning Prompt

Below are the prompt templates used for generating clip caption. The video clips are processed by the VLM at one frame per second (1 fps).

Subsequently, we always ask the model to summarize its produced output into a few sentences.

A.1.1 Clip Captioning Prompt with Character Identification

For the prompt with character identification, we additionally provide the transcripts as input to the model. In that case, the model can internally match every line from the transcripts to the audio of the video in order to infer the names of the characters appearing in every frame.

> <VIDEO CLIP>
>
> Here are the transcripts for the corresponding video:
> <CLIP TRANSCRIPTS>
>
> Describe what is happening in the video in all the details.
> Explicitly state the names of the characters in your description when possible.

A.1.2 Clip Captioning Prompt without Character Identification

The prompt without character identification is used essentially for the ablation study (see Section 5.7). Also, since Qwen2-VL does not have access to the audio, we do not perform character identification with this model and always use the prompt below.

> <VIDEO CLIP>
>
> Describe what is happening in the video in all the details.

A.2 Summarization Prompts

We provide here the prompts we used for generating summaries in our experiments. We write in red the part of the prompt that varies across the tested models.

A.2.1 Screenplay Summarization Pipeline

> Summarize every single existing subplot from the above dialogue. For each subplot, include throughout your summary any important visual detail or information about character actions, interactions, scene location that you may find in the Video Captions.

A.2.2 VLog

> Summarize every single existing subplot from the above dialogue. For each subplot, include throughout your summary any important visual detail or information about character actions, interactions, scene location.

A.2.3 Gemini 1.5 Pro and Qwen2-VL (video)

Summarize every single existing subplot from the above dialogue. For each subplot, include throughout your summary any important visual detail or information about character actions, interactions, scene location that you may pick up from the video frames and provided images.

A.2.4 Simple Summarization Prompt

Summarize every single existing subplot from the above dialogue.Your summary should be very complete.

A.3 Multimodal Summary Evaluation Prompts

A.3.1 Fact Identification

We provide below the few-shot examples and prompts we used for extracting the facts from one groundtruth summary sentence.

Please break down the following sentence into independent facts:
Katie went to Al's diner and reacted to a 'Closed' sign on the door.
- Katie went to Al's diner.
- Katie reacted to a 'Closed' sign on the door.

Please break down the following sentence into independent facts:
Simon ushered Lily in, and she spied a romantic candlelit table just for two.
- Simon ushered Lily in
- Lily spied a romantic candlelit table just for two.

Please break down the following sentence into independent facts:
Luke shouted at his lover that the awards were all for him, but Noah shoved the award into Luke's hands and went inside.
- Luke shouted at his lover that the awards were all for him.
- Noah shoved the award into Luke's hands and went inside.

Please break down the following sentence into independent facts:
Bridget suggests that perhaps Eric can help them
- Bridget suggests that perhaps Eric can help them

Please break down the following sentence into independent facts:
At work at the diner, Simon called Metro to make a dinner reservation, as a customer requests him for service and gives him a hard time making fun of him.
- Simon is at work at the diner.
- Simon called Metro to make a dinner reservation.
- A customer requests Simon for service.
- A customer gives Simon a hard time.
- A customer makes fun of Simon.

Please break down the following sentence into independent facts:
Noah and Jack became furious and accused Luke of taking over their work in order to control them.
- Noah and Jack became furious at Luke.

- Noah and Jack accused Luke of taking over their work in order to control them.

Please break down the following sentence into independent facts:
At Marone, Taylor pays Nick a visit.
- Taylor and Nick are at Marone.
- Taylor pays Nick a visit.

Please break down the following sentence into independent facts:
<INPUT FACT>

A.3.2 Visual Fact Classification

By definition, a visual fact is a fact that cannot be inferred from the transcripts alone. We therefore separate visual from textual facts by asking the LLM the following question.

Fact Evaluation against Transcripts

<TRANSCRIPTS>

Here is a Fact: <INPUT FACT>.
Suppose you are given only the above Transcripts. You do not have access to the Fact. You want to explain to someone everything that happened in the full transcripts above. Do you think that your explanation will contain the given Fact?

Answer by True or False. Justify your answer.

In addition, we manually hand-label a set of facts as either `Visual` or `Textual` (as below) and prompt the LLM to classify each given fact in a few-shot setting using the examples below.

Few-shot Visual Fact Classification

You are given a list of Facts extracted from a movie.
In what follows, your mission is to tell whether the fact is related to what is seen on the screen or to the conversation between the different characters of the story.
Can the fact be deduced from the conversation between the characters?

Fact: There was a ‘Closed’ sign on the door.
Answer: False

Fact: Simon ushered Lily in.
Answer: False

Fact: Luke shouted at his lover that the awards were all for him.
Answer: True

Fact: Noah shoved the award into Luke’s hands and went inside.
Answer: False

Fact: Molly felt that something was not quite right about the situation.
Answer: True

Fact: Simon is at work at Al’s diner.

Answer: True

Fact: The client's name is Laura and she is frustrated with the service.
Answer: True

Fact: Holden notices a look in Molly's eyes that indicates that Molly does not believe Meg.
Answer: True

Fact: The TV show is aimed at mothers with children.
Answer: True

Fact: Simon is having a hard time dealing with the customer.
Answer: True

Fact: There was a noticeable change in Meg's condition.
Answer: True

Fact: Tim was making fun of John.
Answer: True

Fact: Paul requests the teacher for service.
Answer: True

Fact: Lucinda asked a couple of direct questions about Lily being pregnant.
Answer: True

Fact: Noah accused Luke of taking over his work in order to control him.
Answer: True

Fact: <INPUT FACT>
Answer:

A.3.3 Fact evaluation

We provide below the prompt for evaluating the recall for any generated summary. We use the following question for testing whether a fact is supported by the groundtruth summary.

<TRANSCRIPTS>

Is the Input supported by the above summary?
Input: <INPUT FACT>.

Answer by True or False. Justify your answer.

B Human Validations

We conduct a human evaluation of various components of our summarization pipeline as well as the MFactSum metric. All evaluations are performed by a single annotator on the same 4 episodes, sampled from the SummScreen3D test set. Our human evaluation is statistically significant as each episode comprises numerous instances of what is evaluated (e.g. facts, clips, ...).

- Episode 1 is the episode from *As the World Turns* aired on 01-05-2010.
- Episode 2 is the episode from *Guiding Light* aired on 01-25-2005.
- Episode 3 is the episode from *the Bold and the Beautiful* aired on 05-29-2006.
- Episode 4 is the episode from *the Bold and the Beautiful* aired on 06-12-2006.

B.1 Human Validation of the Visual Fact Classification

We include here the results of our human evaluation for the visual fact classification proposed in Section 4.2.

As described in Section 4.2, we extract all the groundtruth summary facts and automatically classify them as either `Visual` (video related) or `Textual` (transcript related) using our metric. We also ask our human evaluator to independently label each fact as `Visual` or `Textual`.

We then study how the classification performed by our metric compares against those human annotations. Over the 4 episodes (573 facts) on which we perform human evaluation, our metric performs visual fact classification with an accuracy of 86.3%.

	Episode 1	Episode 2	Episode 3	Episode 4	Average/Total
`Visual` Fact Precision (%)	90.0	100	86.67	93.65	92.58
`Textual` Fact Precision (%)	84.92	89.47	81.19	87.60	85.80
`Visual` Fact Recall (%)	61.04	50.0	40.63	50.0	50.42
`Textual` Fact Recall (%)	97.13	100	97.62	99.07	98.46
`Visual` Fact Count	77	12	32	30	151
Fact Count	257	63	116	137	573

Table 3: **Human evaluation of the visual fact classification.** We provide the estimated Precision and Recall for both `Visual` and `Textual` facts for each episode and aggregated over all episodes. We also report the total number of facts per episode as well as `Visual` facts found by the human evaluator.

We further study the precision and recall for each class in Table 3. We are able to achieve a high average precision on both `Visual` (92.58%) and `Textual` (85.80%) facts. This means our evaluation metric is robust as we are able to clearly distinguish between `Visual` and `Textual` facts.

While the precision for `Visual` facts is very high (92.58%), the corresponding recall is lower (50.42%). This shows the task difficulty. In the meantime, leveraging the whole video in order to classify `Visual` facts is currently out of reach due to the poor performance of VLMs for long video understanding (see Limitations). By keeping a high precision for `Visual` facts, at the expense of a lower recall, we ensure the robustness of MFactSum by not introducing noise in later stages of the metric computation.

B.2 Human Validation of the Clip Selection Strategy

We provide here a human evaluation for the clip selection strategy proposed in Section 3.1.

We study how our clip selection strategy compares to the random baseline. For the random baseline, we randomly select clips from within the whole episode video. For fairness, we also make sure that we select the same number of clips as in our own clip selection strategy and that the total sampled duration is the same.

For each visual fact from the groundtruth summary, we ask a human evaluator to manually check whether the fact is supported by one of the video clips retrieved by either the random baseline or our clip selection strategy.

% of visual facts in clips	Episode 1	Episode 2	Episode 3	Episode 4	Average/Total
our clip selection	77.6	66.7	53.3	73.3	67.7
random baseline	49.0	16.7	40.0	40.0	36.4
Visual Fact Count	77	12	32	30	151

Table 4: **Human evaluation of our clip selection strategy against the random baseline.** We provide the ratio of visual facts retrieved by either clip selection method. We also report the total number of `Visual` facts in each episode groundtruth summary.

The results of the human evaluation are given in Table 4. Over the 4 episodes (151 visual facts) used for the human validation, our clip selection strategy is found to retrieve about 1.9 times more visual information than the random baseline.

B.3 Human Validation of the Character Identification

We include here the results of our human evaluation for the character identification strategy proposed in Section 3.2.

As described in Sections 3.1 and 3.2, we begin by applying our clip selection algorithm to each episode. For every selected clip, we generate a caption that includes predicted character names. We then ask a human evaluator to name all the characters appearing in each video clip and compare to those found in our generated captions.

Following (Han et al., 2024), we report the IoU score to assess the performance of our method to correctly identify characters in the generated clip captions. The IoU score is given by:

$$\text{IoU} = \frac{|\mathbf{E}_{\text{pred}} \cap \mathbf{E}_{\text{human}}|}{|\mathbf{E}_{\text{pred}} \cup \mathbf{E}_{\text{human}}|}$$

where $|\mathbf{E}_{\text{pred}} \cap \mathbf{E}_{\text{human}}|$ is the number of distinct characters correctly identified in the generated clip caption and $|\mathbf{E}_{\text{pred}} \cup \mathbf{E}_{\text{human}}|$ is the total number of distinct characters found by either the human evaluator or in the generated clip caption.

	Episode 1	Episode 2	Episode 3	Episode 4	Average/Total
IoU score (%)	81.89	74.06	69.71	83.89	77.39
Clips Count	69	50	25	30	174

Table 5: **Human evaluation of the character identification strategy.** We provide the average IoU score for each episode and aggregated over all episodes. We also report the total number of clips for each episode on which human evaluation is performed.

Over the 4 episodes (174 video clips), we show in Table 5 that our generated clip captions share 77.4% of characters in common with the original video according to our human evaluator. This makes our character identification prompting a reasonable choice for the studied dataset. We found those results to be on par with other strategies used for movies or TV shows as proposed in (Han et al., 2024) and (Xie et al., 2024) in which they respectively found a score of 70.8% and 75.8% based on 4 movies from the MovieNet dataset (Huang et al., 2020).

C Summary Imbalance and Textual Bias in Finetuned Models

Table 6 shows the gold summary for one episode from the TV show *The Bold and the Beautiful* aired on May 5, 2006. After watching the whole episode, we manually highlight in green all the visual content present in the groundtruth summary. We notice that visual information, while essential to understanding the episode, is present in only a minority of the groundtruth summary.

The summary produced by Modular-Kosmos (see Table 7) completely ignores all the visual content from the episode only focusing on relating the conversation between the characters. As shown in

Table 7, most of the summary sentences generated by the model are indeed limited to "[Someone] says", "[Someone] tells" or "[Someone] asks". This may be due to the limited presence of visual information in training summaries, leading finetuned models to focus primarily on dialogue understanding of an episode.

The gold Summary of *The Bold and the Beautiful* episode (aired 05-05-06)

Ridge continues to beg Brooke to reconsider her decision to leave Forrester as Stephanie continues to voice her opinion. At Marone, Taylor pays Nick a visit. Nick is still angry about what Taylor implied when she disclosed that Brooke and Ridge slept together. Taylor tries to apologize and asks if things are all right between Nick and Brooke. Nick tells her that everything is fine and Brooke is quitting her job at Forrester. Taylor is unconvinced that Brooke will be able to let go of either Forrester or Ridge ! Brooke tells Ridge that she cannot fight with Stephanie any longer and that her future is with Nick. After kissing Ridge and saying that a part of her will always love him, she takes her things and leaves. Bridget and Dante are at home discussing Stephanie's interference in the custody of Dino. Bridget suggests that perhaps Eric can help them. Dante worries about what losing his job would do to his work visa. Bridget convinces him that because they all love Dino, they should be able to work something out. After some wine, Bridget reveals that she is ready to make love with Dante. As the two were in bed, Dante stops and reaches in the bedside drawer and presents Bridget with an engagement ring and pops the question. Brooke goes to see Nick at his office and tells him that she has left Forrester. Nick is pleased, although Brooke confesses that she hurt Ridge badly by walking out. Nick whisks her off to the Marone jet for a surprise getaway! At Forrester, Ridge angrily accuses Stephanie of causing all his problems with Brooke. Stephanie is stunned as Ridge bashes her with a vengeance and then clutching his chest, collapses to the floor!

Table 6: Gold summary from the SummScreen3D dataset for *The Bold and the Beautiful* episode (aired 05-05-06). Several summaries exist for this episode from different fansites. This summary is from TVMegaSite.

Modular-Kosmos Summary of *The Bold and the Beautiful* episode (aired 05-05-06)

Bridget and Dante plan to get married in Italy, but Bridget wants to spend more time with Dante. Stephanie tells Bridget that she wants to fire Dante and send him to Italy. Ridge tells Stephanie that he wants Brooke out of Forrester Creations. Brooke tells Nick that she is through fighting, she is moving to Paris. Nick tells her that she has to move out of the office. Nick says that he will not force her to work in Paris, but she says that she needs to move in. Ridge says that Nick is her future, and she has no choice but to leave. Stephanie says that Ridge needs her, but he does not want to accept her. He says that they are better off without her. She says that is not what she wants. She tells him that she does not know how he feels. She does not understand how he could do this to Brooke. He tells her to move on with her life. Taylor apologizes to Nick for suggesting that Ridge and Brooke slept together. Nick tells Taylor that he doesn't let it affect his relationship with Brooke. Brooke tells Nick that she needs something to cheer her up, and Nick tells her that they are on their way to a tropical island. Ridge tries to convince Brooke to stay at Forrester Creations, but she tells him that she is leaving anyway. Taylor tells Ridge that she does not want to believe that he and Brooke have feelings for each other, but Ridge tells Taylor not to give up on her dreams. Nick and Brooke plan to take off on their trip, but Nick tells Brooke that he has a surprise for her, and that they will be going to a beautiful beach and a fruity drink. Brooke is very happy about Nick's trip plans, but when she asks where they are going, Nick reveals that he wants to take her to a romantic beach.

Table 7: The summary generated by Modular-Kosmos model for *The Bold and the Beautiful* episode (aired 05-05-06).

D Additional Metrics

We report in Tables 8 and 9 additional results on METEOR and BertScore metrics. Consistently with our experiments on ROUGE (Section 5.6), we always take the maximum score against all references. We respectively leverage the `meteor_score` function from `nltk.translate` and the `bert_score` python package.

Similar to what we observed with ROUGE (see Section 5.6), we found that both metrics are not able to tell the difference between a multimodal and a monomodal summary. Indeed, both METEOR and BertScore in Table 8 were found to be even higher when videos were removed from our screenplay summarization pipeline, regardless of the base model used (Gemini 1.5 Pro or Qwen2-VL-72B).

	METEOR	bert-prec	bert-rec	bert-f1	avg-len
multimodal baselines					
Modular-Kosmos (Mahon and Lapata, 2024)	31.45	**83.52**	**85.03**	**84.25**	314.0
VLog	21.09	80.35	82.31	81.30	314.0
Qwen2-VL-72B (no video)	29.46	81.11	84.59	82.80	718.5
Qwen2-VL-72B (video)*	21.53	80.04	82.51	81.25	889.0
Screenplay Summary (Qwen2-VL-72B)	20.79	79.48	82.42	80.92	749.5
Gemini 1.5 Pro (no video)	31.83	82.52	84.81	83.6	573.94
Gemini 1.5 Pro (video)*	**33.38**	81.75	84.89	83.28	688.3
Screenplay Summary (Gemini 1.5 Pro)	27.97	81.68	83.99	82.81	601.1

Table 8: **METEOR and BertScore Evaluation Results on SummScreen3D.** We respectively denote BertScore precision, recall and F1 score by `bert-prec`, `bert-rec` and `bert-f1`. The average summary word count is denoted by `avg-len`. Best results are in **bold**. * indicates the VLM is prompted on the full video and transcripts in an end-to-end fashion using the maximum number of frames allowed by the API.

	METEOR	bert-prec	bert-rec	bert-f1	avg-len
w/o handcrafted prompt	**27.97**	**81.99**	83.85	**82.90**	609.9
w/o character ident.	27.76	81.53	83.88	82.68	576.5
Screenplay Summary (Gemini 1.5 Pro)	**27.97**	81.68	**83.99**	82.81	601.1

Table 9: **Ablation results on METEOR and BertScore for our screenplay summarization pipeline.** We respectively denote BertScore precision, recall and F1 score by `bert-prec`, `bert-rec` and `bert-f1`. The average summary word count is denoted by `avg-len`. Best results are in **bold**.

E Models comparison based on visual recall ability

We consider the episode from the TV show Guiding Light aired on January 25, 2005, for which we provide the groundtruth summary in Table 10. After watching the whole episode, we compare the summaries produced by our baselines (Section 5.4) to our own screenplay summary produced with Gemini 1.5 Pro as the base model (see Table 11).

In all the summaries, we manually highlight in green the visual information that is supported by the groundtruth, in yellow, the visual information that is present in the episode but not mentioned in the groundtruth and in red all the visual information that is hallucinated by the model.

For the chosen example, we notice our screenplay summaries are able to recall (in green) more visual information than any other tested baselines including Gemini 1.5 Pro prompted end-to-end on the full videos and transcripts (see Table 12).

We also report the visual recall that we automatically compute using Gemini 1.5 Flash for all models' summaries of the chosen episode. The results confirm our qualitative analysis: our screenplay summary successfully retrieves more visual facts than any other tested baseline.

Groundtruth Summary of *Guiding Light* episode (aired 01-25-05)
Josh and Reva have gone out to dinner to try to get away from the ”situation” at home. But Reva can’t help herself, and tries to call Jonathan to check in on him. She doesn’t trust the nurse they hired to take care of him while they were out. Her maternal instincts were right! Cassie talked the nurse into trading places with her so she could have Jonathan to herself. She confronts Jonathan about flowers which were sent to her wedding suite, but Jonathan denies everything. Dopey from pain killers and beer, he falls asleep. Cassie takes a close look at his pain killers and just as she pours a few in her hand, Josh and Reva walk in and see her. Reva demands to know what she has done to her son, but Cassie says she has done nothing. Reva is very angry and Josh and Cassie are angry with her. Josh tries to tell her that he is sick of the anguish Jonathan is causing the family. Cassie says her sister is ”gone.” When Cassie leaves to go home, Josh follows her. He thinks this is a family issue and they will work it out. But Cassie reminds him the problem is Reva thinking of Jonathan as family, too. Indoors, Reva tells Jonathan things will work out, as they always do. Through some detective work, Harley has found Beth at a spa. She tries to talk with her, but Beth is hysterical. Harley assumes it is because Beth wants to confess that she killed Phillip. But instead, Beth admits the night Phillip was killed, she was going to beg to go with him! Harley is furious when Beth goes on to tell her that she wanted to go with all the children, including Harley’s, and that no one would ever have seen the children again. Beth is shocked at herself that she could have considered such a plan, but Harley is crestfallen when she realizes this means she may have killed Phillip. She is inconsolable. But. . . back in Springfield, Gus is chatting with Lizzie. She tells him her mother was with her every minute of the evening Phillip was killed. This directly contradicts what Beth has just told Harley. Does Lizzie believe her mother killed her father? Did Lizzie kill her father? Unfortunately, Alan is very angry at Gus and has followed him to Elizabeth & Company to have words. Alan overhears a cell phone conversation between Gus and Harley, including when Gus tells Harley the judge would be very angry if he found out she had crossed state lines.

Table 10: Groundtruth summary for the *Guiding Light* episode (aired 01-25-05).

Screenplay Summary of *Guiding Light* episode (aired 01-25-05)

1. Reva and Josh's Relationship Struggles Over Jonathan: This subplot centers around the tension between Reva and Josh caused by Jonathan's manipulative behavior. Reva consistently defends Jonathan, even against accusations of drugging Cassie the night before her wedding and other problematic actions. Josh is increasingly frustrated with Reva's blind loyalty to her son. They have a tense conversation in their living room, where Reva accuses Josh of taking Cassie's side. The argument escalates, with Josh expressing concern about Reva's enabling behavior and Reva feeling isolated and misunderstood. Visual cues like Reva offering a handshake before exiting the room, pausing mid-departure, and Josh pursuing Cassie onto a dimly lit porch highlight the growing distance between them.

2. Cassie's Quest for Answers and Revenge against Jonathan: Cassie is deeply disturbed by Jonathan's actions, particularly the implied drugging incident the night before her wedding. She confronts him multiple times, demanding the truth. Jonathan, sitting in a wheelchair, is evasive and mocking. Cassie's desperation is evident as she initially considers harming him, holding a bottle of pills. She later vents her frustration to Reva and Josh, revealing the depth of her anger and hurt. The visual of Cassie shaking pills into her hand while Jonathan dozes emphasizes her contemplation of revenge.

3. Gus's Determination to Clear Harley's Name: Gus firmly believes in Harley's innocence in Phillip's murder and is tirelessly pursuing other leads. He clashes with Alan, who believes Harley is guilty and sees Gus's efforts as a betrayal of the family. Gus questions Lizzie about Beth's whereabouts the night of the murder, hoping to uncover information that could exonerate Harley. The tension between Gus and Alan plays out in an office setting, near a couch, bookshelf, and lamp. Gus's desperation is highlighted by his abduction of Lizzie, covering her mouth and forcing her towards an exit.

4. Beth and Harley's Shared Secret: Beth confesses to Harley her plan to leave the country with Phillip and their children, a plan she abandoned upon arriving at Company. Harley, in turn, reveals her own turmoil and uncertainty about her actions that night, suggesting she may have been the one who shot Phillip. The two women share a moment of vulnerability and mutual understanding in a spa-like room with candles and a potted plant, promising to keep each other's secrets. Visuals like Beth sitting on the floor sobbing, Harley kneeling beside her in a dark red jacket, and Beth's reflection in the mirror underscore their emotional distress.

5. The Mystery of Phillip's Murder: The overarching plot revolves around the unsolved murder of Phillip. While Harley is the prime suspect, Gus's investigation suggests other possibilities. Lizzie provides an alibi for Beth, claiming she was with her mother the entire night. The dialogue heavily focuses on the events of that night, with various characters recalling their actions and whereabouts. The mystery is fueled by conflicting accounts, withheld information, and the emotional turmoil of the characters involved.

Table 11: The screenplay summary for the *Guiding Light* episode (aired 01-25-05). Our screenplay summary is able to recall 67% of the visual information from the groundtruth summmary.

Gemini Summary of *Guiding Light* episode (aired 01-25-05)

1. Reva and Josh are on a date. Josh is trying to have a good time with Reva, but she is distracted, constantly worried about her son Jonathan and wanting to call home. It is revealed later that Jonathan is under the care of a nurse due to his being in a wheelchair. Reva doesn't trust the nurse with the medication, however, Josh tries to convince her that Jonathan is fine.
2. Cassie comes to Jonathan's house, finds him alone and sitting in a wheelchair by the fireplace, with some beer bottles by his side. Cassie wants to know what happened between her and Jonathan on the night before the wedding. She gives him a card that he sent her; it read, "Stork doesn't judge. A baby is a blessing even if it is conceived the night before the wedding." Jonathan plays dumb, but eventually, Cassie reveals that Jonathan drugged her and she doesn't remember what happened, and whether they slept together or not.
3. Gus is talking to Alan. Gus wants to find Beth because he thinks that she might know something about Phillip's death, and he wants to clear Harley's name. Alan, however, believes that Beth doesn't know anything. He wants Gus removed from the company because ever since Phillip's death, Gus has not been a good Spaulding. Gus remembers that he lied for Alan and Alexandra regarding the drugs, which cost him his wife. As far as Alan is concerned, Gus is siding with Harley and this will cost Beth her kids. Gus reminds Alan that Harley is also a mother and her son is Alan's grandson.
4. Harley is talking to Beth. Beth admits to Harley that Phillip was going to take her children away and she had to stop him. She was going to ask Phillip to take her and the kids away. Harley thinks that was a smart idea to make Phillip reveal where the kids were. Beth confesses that she was going to leave with Phillip. Harley wants to know how she could rob another mother of her kids. Beth says she should ask Harley, and what did she do when she was planning to leave? Harley says she doesn't remember if she killed Phillip. Beth says that the only weapon she had was history and she knew that Phillip would not believe her that night, so she changed her mind about leaving.
5. Gus meets with Lizzie in the house and he asks her where her mother is. He explains he needs to talk to her, ask her a few questions about the night Phillip was killed. Lizzie doesn't like Gus and thinks that he doesn't believe Harley is innocent. Gus claims that the evidence against Harley is circumstantial. Lizzie still refuses to tell Gus where her mother is.
6. Reva goes home and finds Cassie there. She is angry. Cassie is accused of wanting to kill Jonathan, because Jonathan is drunk. Cassie and Josh reveal to Reva what Jonathan did to Cassie on the night before the wedding. Reva finds out that Jonathan drugged and slept with Cassie. Reva and Josh have an argument over Jonathan's behavior, and about how Reva always takes Jonathan's side.
7. Gus questions Lizzie about the night her father died and where she and her mother were. Lizzie claims that she was with her mother the entire time and that she didn't kill Phillip. Gus suspects that Lizzie is protecting her mother by giving her an alibi.
8. Harley tells Beth to keep their conversation about the night Phillip died a secret between them. Beth owes Harley one. Gus calls Harley asking if she found Beth. Harley says that Beth was innocent. Gus thinks they can find another lead now, but Harley says she cannot do it anymore. During their conversation, Harley takes something out of her handbag and puts it back. She drops her handbag and something falls out, but she doesn't pick it up. She seems worried and scared.

Table 12: Gemini's generated summary for the *Guiding Light* episode (aired 01-25-05). Gemini's summary is able to recall 17% of the visual information from the groundtruth summary.

VLog Summary of *Guiding Light* episode (aired 01-25-05)

Based on the dialogue and visual details provided in the video, there are several subplots that can be identified:

1. The first subplot involves a man in a gray jacket and white shirt being questioned about the night his father died. He is asked about his mother's whereabouts and what she was wearing that night. The man insists his mother is innocent and refuses to provide further information, leading to suspicion from the questioner.

2. The second subplot features a woman in a white robe and another woman discussing a shared secret. They agree not to judge each other for their actions that night and decide to keep the information between themselves.

3. The third subplot revolves around a man talking on a cell phone while wearing a jacket. He informs someone named Gus that a woman named Beth is innocent and did not see Phillip on the night he was killed. The man expresses reluctance to point fingers at other suspects and seems to be struggling with his own actions.

4. The fourth subplot shows a woman holding a leaf up to her face and expressing a desire for justice. She threatens someone named Phillip and vows to make him pay for his actions.

5. The fifth subplot involves a woman in a blue shirt holding a remote control and interacting with a nurse. She expresses concern about her son's medication and decides to leave him in the nurse's care while she steps away for a moment.

6. The final subplot features a man in a black jacket talking on a cell phone and confronting someone named Alan. The man is determined to clear Harley's name and warns Alan not to interfere with his efforts.

These subplots involve a mix of mystery, secrets, suspicions, and confrontations, adding depth to the overall narrative of the video.

Table 13: VLog's generated summary for the *Guiding Light* episode (aired 01-25-05). VLog's summary is able to recall 0% of the visual information from the groundtruth summary.

Modular-Kosmos Summary of *Guiding Light* episode (aired 01-25-05)

At the hospital, Reva finds out that Jonathan has had his full dose of her fertility drugs and that he has no memory of the night of the wedding. At Company, Alan tells Gus that he needs to remove Gus from the company. Gus tells Alan that he is trying to clear Harley's name. Lizzie goes to the police station and tells them that Harley did not kill her father. Gus tries to convince her that Harley is innocent. Alan tells Alan to get rid of Gus. Alan asks Gus if he knows something about the night Phillip died. Gus says that he knows nothing. Alan says that Alan needs to have Gus removed from Company. At the farm, Cassie asks Jonathan if he remembers the night before the wedding, but he says he doesn't remember. She asks him if they did the deed. He tells her that he does, but she says that she doesn't have to tell him. She tells him that she is on fertility drugs. He Reva tells Cassie that she's been drinking all night long and taking painkillers. Reva asks Cassie to give her the pills. Cassie tells Reva that the night before her wedding she was staying at Jonathan's place. Josh tells Jonathan that Reva threw him down the stairs. Jonathan tells Josh that he doesn't know what's going on. Lizzie tells Gus that Phillip was shot dead. Gus tells her that she needs to tell him what happened the night Phillip died. Harley tells Beth that she doesn't want her children to grow up without Phillip. Beth tells Harley that she didn't kill Phillip. Harley is upset and tells Beth to tell her what happened to Phillip. Gus asks her if she wants to talk to him. Harley says she's tired and wants to get out of the car. Gus takes her to the police station, where she tells him that Phillip is dead. Harley asks if she's going to tell Gus what happened. Gus

Table 14: Modular-Kosmos's generated summary for the *Guiding Light* episode (aired 01-25-05). Modular-Kosmos's summary is able to recall 17% of the visual information from the groundtruth summary.