Hierarchical3D Adapters for Long Video-to-text Summarization

Pinelopi Papalampidi* **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
p.papalampidi@sms.ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

In this paper, we focus on video-to-text summarization and investigate how to best utilize multimodal information for summarizing long inputs (e.g., an hour-long TV show) into long outputs (e.g., a multi-sentence summary). We extend SummScreen (Chen et al., 2022), a dialogue summarization dataset consisting of transcripts of TV episodes with reference summaries, and create a multimodal variant by collecting corresponding full-length videos. We incorporate multimodal information into a pre-trained textual summarizer efficiently using adapter modules augmented with a hierarchical structure while tuning only 3.8% of model parameters. Our experiments demonstrate that multimodal adapters outperform more memory-heavy and fully fine-tuned textual summarization methods.

1 Introduction

What happens in the very last episode of "Friends"? Anyone who has seen this episode can summarize its key moments: Ross confesses his love for Rachel, they decide to resume their relationship, while Monica and Chandler adopt twins and move to the suburbs. TV viewers can naturally perform this dialogue summarization task having access to multiple modalities: they not only hear the actors speak but also see their expressions, actions, and whereabouts on screen.

Despite recent advances in summarization (Nallapati et al., 2016; See et al., 2017; Liu and Lapata, 2019b) and increasing interest in different types of dialogue summarization, e.g., meeting transcripts (Gliwa et al., 2019; Zhong et al., 2021) or screenplays (Chen et al., 2022), the contribution of modalities other than text remains relatively understudied. This is not entirely surprising given the challenges associated with the multimodal summarization task illustrated above (e.g., produce a written summary of a TV episode). Firstly, the input is long, it cannot fit into standard sequence-to-sequence architectures, and the different modalities have to be somehow combined; secondly, the output is also long, summaries consist of multiple sentences and rich vocabulary; and thirdly, it involves complex inference over long-range dependencies between events and characters and common sense reasoning. At the same time, creating large-scale multimodal datasets with long videos and aligned textual data is challenging and time consuming, limiting the research conducted in this domain.

Previous work on video-to-video summarization identifies highlights from YouTube videos, TV shows, or movies (Song et al., 2015; Gygli et al., 2014; De Avila et al., 2011; Papalampidi et al., 2021b). However, in most cases, either the videos are short or the datasets are small with a few hundred examples. There is also limited work on video-to-text summarization. We are only aware of one large-scale multimodal dataset for this task, namely How2 (Sanabria et al., 2018), which again contains short videos (i.e., 2–3 minutes long) with simple semantics, and short, single-sentence summaries.

In this paper, we focus on video-to-text summarization and investigate how to best utilize multimodal information for condensing long inputs (e.g., an hour-long TV show) into long outputs (e.g., a multi-sentence summary). We create a multimodal variant of SummScreen (Chen et al., 2022), a recently released dataset comprising of transcripts of TV episodes and their summaries. We collect full-length videos for 4,575 episodes and multiple reference summaries. We build our model on top of a pre-trained sequence-to-sequence architecture (i.e., BART; Lewis et al. 2020) fine-tuned on summarization and capable of generating fluent long text. We convert its textual encoder to a multimodal one by adding and tuning adapter layers (Rebuffi et al., 2017; Houlsby et al., 2019),

*Now at DeepMind.

Findings of the Association for Computational Linguistics: EACL 2023, pages 1297–1320

	Modality	Input	Output	Datasets
text-to-text	text	short	short	XSum (Narayan et al., 2018), CNN-DailyMail (Nallapati et al., 2016), NYT (Durrett et al., 2016), Gigaword (Napoles et al., 2012)
	text	long	long	SamSum (Gliwa et al., 2019), QMSum (Zhong et al., 2021), SummScreen (Chen et al., 2022)
video-to-video	vision	short	short	OVP (De Avila et al., 2011), YouTube (De Avila et al., 2011), SumMe (Gygli et al., 2014)
	vision/text	short	short	TVSum (Song et al., 2015)
	vision/text(/audio)	long	long	LoL (Fu et al., 2017) TRIPOD+ (Papalampidi et al., 2021b)
video-to-text	vision	long	short	TACoS (Rohrbach et al., 2014)
	vision/text/audio	short	short	How2 (Sanabria et al., 2018)
	vision/text/audio	long	long	SummScreen^{3D}

Table 1: Datasets used for summarization grouped based on the input/output modalities and input/output length. A more detailed comparison and statistics for video-to-text datasets can be found in Appendix A (Table 10).

which only account for 3.8% of model parameters. We also explore strategies for *content selection*, since the input is too long to fit into standard sequence-to-sequence models. Empirical results across evaluation metrics demonstrate that multimodal information yields superior performance over just text, both in terms of content selection and summarization; this is the case even when our adapter model is compared to fully fine-tuned approaches and more memory-heavy architectures (e.g., Longformer; Beltagy et al. 2020) that can process the entire input.

Our contributions can be summarized as follows: (1) we augment SummScreen (Chen et al., 2022) with multimodal information, providing videos aligned with transcripts and summaries; to the best of our knowledge, this constitutes the largest available resource for long video multimodal summarization; (2) we propose a *parameter efficient* approach to augment a pre-trained textual summarizer with multimodal information; and (3) explore different methods for identifying salient moments in a long video and show that multimodal information also improves content selection.

2 Related Work

Video Summarization Much previous work has focused on text-to-text or video-to-video summarization. We provide a comprehensive categorization of existing datasets according to input/output length and modality in Table 1. *Multimodal abstractive summarization* (video-to-text) has attracted less attention, mainly due to the difficulty of collecting large-scale datasets. How2 (Sanabria et al., 2018) is the only publicly available benchmark for this task, it includes short instructional videos with textual transcripts and one-sentence summaries. We generate multiple-sentence summaries from long videos and their transcripts. While previous approaches have focused on various modality fusion methods with small RNN-based models (Palaskar et al., 2019), we take advantage of large pre-trained LMs (Lewis et al., 2020; Raffel et al., 2020; Radford et al., 2019) for generating fluent text summaries.

Recent years have also witnessed increasing interest in multimodal video captioning, a task related to multimodal summarization, which aims to generate one-sentence descriptions for localized events in short videos (Xu et al., 2016; Rohrbach et al., 2017; Zhou et al., 2018; Lei et al., 2020b). Existing methods employ strong language-and-vision encoders with massive pre-training (Li et al., 2020; Luo et al., 2020; Xu et al., 2021; Lei et al., 2020a; Li et al., 2021), while the decoder is typically shallow and under-trained.

Realizing the importance of large LMs for generation, recent work has focused on how to efficiently render pre-trained LMs multimodal. Notably, Tsimpoukelli et al. (2021) convert a pre-trained LM into an image captioning model, by giving images as prompts and training only a vision encoder. Yu et al. (2021) summarize How2 videos by augmenting BART with visual information via *a new cross-attention block* added to *every encoder layer*. However, their approach adds a very large number of *new parameters* and requires full fine-tuning, which leads to overfitting in our case when the dataset size is small.

Dialogue Summarization In the context of text-to-text generation, dialogue summarization is challenging due to the difficulty of fitting very long input into pre-trained sequence-to-sequence models. Longformer (Beltagy et al., 2020) alleviates this by employing local self-attention in combination

Episodes		4,575
Input (transcript + video + audio)		
Shots		1,048,024
Shots/episode		193.64 (109.09)
Utterances/episode		322.76 (116.52)
Tokens/episode		5720.55 (2223.38)
Output (summaries)		
Summaries/episode	1.53	(0.79)
TVMegaSite/#tokens	4,280	395.69 (275.84)
YouTube/#tokens	334	136.22 (45.12)
IMDb/#tokens	946	111.21 (82.18)
tvdb/#tokens	1,454	126.14 (82.14)
Training (unique input-output pairs)		5,199
Validation episodes		296
Testing episodes		296

Table 2: SummScreen^{3D} statistics. For summaries, we show their provenance, number of summaries per site (second column), and mean number of tokens per summary; standard deviations are shown in parentheses.

with global tokens for reducing the computational overhead. Despite recent attempts to make self-attention more efficient (Kitaev et al., 2020; Tay et al., 2020; Zaheer et al., 2020), it is still unclear whether it has an advantage over content selection with a full-attention mechanism (Zhang et al., 2021; Shaham et al., 2022) for long dialogue summarization. Zhong et al. (2022) incorporate dialogue-specific objectives for pre-training summarization models, while Zhang et al. (2022) hierarchically summarize the input chunk-by-chunk.

Parameter-efficient Tuning Fine-tuning is a common approach for transferring pre-trained models to different tasks or domains (Howard and Ruder, 2018). It is customary to fine-tune all the parameters of the pretrained model which, however, becomes prohibitive as model size and number of tasks grow. Recent work has proposed parameter-efficient transfer learning methods which fine-tune only a small number of *additional* parameters. Two popular approaches include *adapter tuning*, where bottleneck layers are added and tuned at every layer of the model (Rebuffi et al., 2017; Houlsby et al., 2019) and *prompt tuning*, where (soft) prompts are prepended as part of the input (Brown et al., 2020; Li and Liang, 2021). In this work, we utilize the former method for adapting a textual summarizer to our multimodal setting and dialogue input format.

3 The SummScreen^{3D} Dataset

SummScreen (Chen et al., 2022) is a long dialogue summarization dataset[1] containing transcripts from TV episodes and human-written abstractive summaries. We extend this dataset to a multimodal setting by also considering the corresponding full-length videos. SummScreen contains two subsets depending on the series genre: SummScreen-FD and SummScreen-TMS. We use the latter subset which mostly covers soap operas from TVMegaSite[2], as it is easier to obtain full-length videos and each series has hundreds of episodes.

For each episode in SummScreen-TMS, we automatically search for the title and release date in Youtube. If there is a match with large duration (indicating that this is a full episode rather than a segment), we download the video and closed captions (CC). Overall, we collected videos for 4,575 episodes from five different shows in SummScreen-TMS.[3] In addition to TVMegaSite summaries (distributed with SummScreen), we further retrieved summaries from YouTube descriptions, IMDb, and tvdb, again using the episode title and release date as search terms. The statistics of our dataset which we call SummScreen^{3D} (3D for language, video, and audio) are in Table 2 and we provide further details in Appendix A. As can be seen, each episode has (on average) multiple references which vary in length (TVMegaSite summaries are longest).

We split SummScreen^{3D} into training, validation, and test sets with the same distribution over different shows per set. We reserved 296 episodes for validation and the same number for testing, and used the rest for training. Since we have multiple reference summaries for some episodes, we increased the size of the training set by adding m episode-summary pairs, matching the same episode with each of its m references. This resulted in 5,199 unique samples for training.

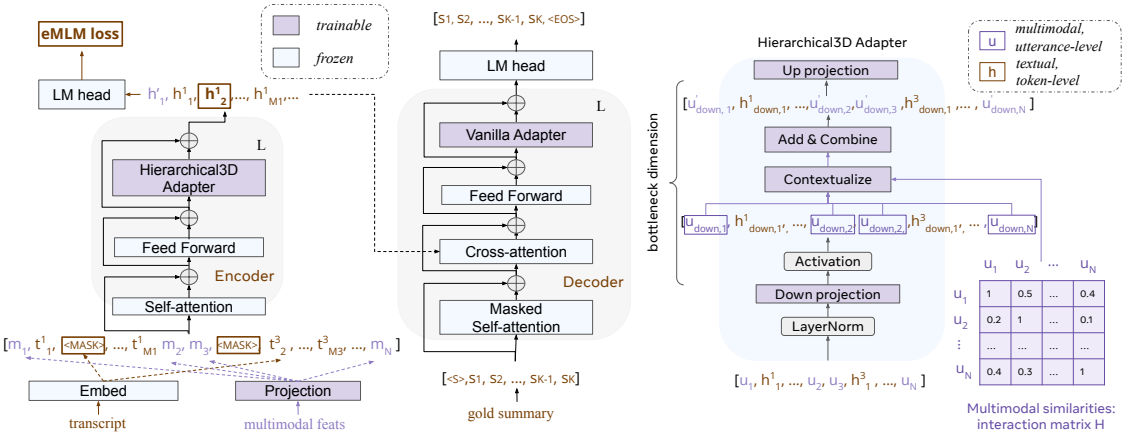
4 Video-to-Text Summarization

Our approach leverages the generation capabilities of large pre-trained sequence-to-sequence models (Lewis et al., 2020; Raffel et al., 2020). As our backbone model, we employ BART-large (Lewis et al., 2020) which has been fine-tuned on CNN-DailyMail (Nallapati et al., 2016; Zhang et al., 2021) and has thus acquired a summarization inductive bias. As TV show transcripts are very long and cannot fit into BART, we select a subset of utterances (i.e., speaker turns) as input via content

[1] https://github.com/mingdachen/SummScreen

[2] http://tvmegasite.net

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



(a) Multimodal augmentation of textual BART.

(b) Hierarchical3D adapter for the encoder layers.

Figure 1: Multimodal augmentation of pre-trained BART. We augment the encoder and decoder layers with adapters which we fine-tune on the target dataset, while the remaining network is frozen. As input, we consider textual tokens and coarse-grained multimodal information which we prepend before each utterance. We also corrupt part of the textual input during training and add an auxiliary MLM loss to the encoder for predicting the corrupted tokens. On the right, we show the hierarchical adapter added to each encoder layer: after down-projecting all representations, we only consider the multimodal ones and further contextualize them via attention. Then, we combine the representations and up-project again to the original model dimension.

selection (see details in Section 5). We transfer this model to our task and domain (i.e., multimodal dialogue summarization), by adding adapter layers (Rebuffi et al., 2017; Houlsby et al., 2019; Sung et al., 2022) in both the encoder and decoder, and tuning them on SummScreen^{3D} while keeping the rest of the network frozen. We briefly discuss below our backbone text-based model and then elaborate on how we incorporate multimodal information.

4.1 Backbone Textual Model

Our summarizer follows a standard sequence-to-sequence Transformer architecture (Vaswani et al., 2017). The encoder maps tokens $[t_1, t_2, \ldots, t_N]$ to a sequence of contextualized representations $[h_1, h_2, \ldots, h_N]$ which are then fed to the decoder for generating the summary. The encoder consists of L stacked layers, each of which has a self-attention block for contextualizing the token representations, followed by a feed-forward network. The decoder has a similar architecture, it additionally contains a *cross-attention* block for identifying relations between the input and currently generated text and makes use of *masked* self-attention to control access to context for each token. The decoder is followed by a linear layer (i.e., Language Model (LM) head) which projects the output representations onto the vocabulary and a final softmax layer. The model is optimized for predicting the next token s_{t+1} in the summary given $[s_0, s_1, \ldots, s_t]$, the context generated so far, and the transcript $[t_1, t_2, \ldots, t_N]$.

4.2 Multimodal Augmentation

Our hypothesis is that adding multimodal information to a textual summarizer (i.e., converting the textual encoder to a multimodal one) will increase the quality of its output summaries. We expect that the video/audio will compensate for important non-verbal information typically absent from the transcript (e.g., who is speaking to whom, who is present in the same room, who is crying or yelling). We further expect multimodal information to make up for the loss of context incurred by content selection. We next describe how we compute multimodal representations for an episode and how we augment BART with these representations.

Multimodal Representations We use *utterances* as the unit of representation for multimodal information. We segment episodes into shots (using PySceneDetect[4]) and map these to utterances in the corresponding transcript. Specifically, we align the closed captions in the video which are time-stamped to the utterances in the transcript using Dynamic Time Warping (DTW; Myers and Rabiner 1981; Papalampidi et al. 2021b). We thus create a one-to-many alignment where an utterance corresponds to one or more shots. For each shot, we extract textual, visual, and audio features (see Appendix B.1 for details), and compute an utterance-level representation for each modality by average pooling over all aligned shots.

Given textual x_i, visual v_i, and audio a_i repre-

[4] https://github.com/Breakthrough/PySceneDetect

sentations for utterance i, we learn a multimodal representation as part of our network:

$$x'_i = f(W_x x_i) \quad v'_i = f(W_v v_i) \quad a'_i = f(W_a a_i)$$
$$m_i = f(W_m[x'_i; v'_i; a'_i]) \tag{1}$$

where $f(\cdot)$ is the ReLU activation function, $[\cdot;\cdot;\cdot]$ denotes concatenation, $W_x \in \mathbb{R}^{d_x \text{x} d_i}, W_v \in \mathbb{R}^{d_v \text{x} d_i}, W_a \in \mathbb{R}^{d_a \text{x} d_i}$, and $W_m \in \mathbb{R}^{3d_i \text{x} d_m}$ are learnable matrices; d_i and d_m are the input and model dimensions with $d_i << d_m$, and m_i is the final multimodal representation corresponding to the i^{th} utterance in the transcript.

Multimodal Encoder In order to integrate utterance-level multimodal representations with BART, we consider a "global utterance token" inspired by the Longformer architecture (Beltagy et al., 2020). We preprocess the input into utterances and prepend a global token <EOS> per utterance as a placeholder for multimodal representations. The encoder thus receives as input sequence $[\mathbf{m_1}, t^1_1, t^1_2, \ldots, t^1_{M_1}, \ldots, \mathbf{m_N}, t^N_1, t^N_2, \ldots, t^N_{M_N}]$ where, "global" representations $\mathbf{m}$ constitute a rich multimodal space (i.e., they are not learned solely from text via local self-attention; Figure 1a).

4.3 Self-supervised Auxiliary Guidance

Our primary loss for training the model described above is the negative log likelihood of predicting the next token in the summary given episode $\mathcal{E}$:

$$L_{LM} = \frac{1}{K} \sum\nolimits_{t \in [1,K]} -\log p(s_t | s < t; \mathcal{E}) \tag{2}$$

We further wish to encourage the model to attend to multimodal information and learn a meaningful projection (Equation (1)). To do this, we corrupt part of the textual input by masking tokens (see bottom left part of Figure 1a) and adding an auxiliary masked language modeling (MLM) loss for the initial training steps only. So as not to disrupt the bias of the decoder, which is already trained on textual summarization, we apply the MLM loss in the outputs of the encoder while the model is trained on the downstream task. Given token-level encoder outputs $[h_1, h_2, \ldots, h_N]$, we copy and re-use the LM head of the decoder in order to project them into the vocabulary (see top left part of Figure 1a). And compute the negative log likelihood only for the set of masked tokens $\mathcal{M}$:

$$\mathcal{L}_{eMLM} = \frac{1}{|\mathcal{M}|} \sum\nolimits_{t \in \mathcal{M}} -\log p(t | h_{t_i \notin \mathcal{M}}) \tag{3}$$

We refer to this loss as encoder-based MLM loss (eMLM; Baziotis et al. 2021). It trains the encoder to reconstruct input text representations while attending to multimodal information. After X initial training steps, we drop the auxiliary loss and stop corrupting the textual input in order for the model to be optimized on summarization. We use a mixture of whole utterance corruption (Zhang et al., 2020a; Zhong et al., 2022) and content word corruption, masking out named entities, nouns, and verbs excluding auxiliaries (see Section 6).

4.4 Hierarchical3D Adapters

We specialize BART for our multimodal summarization task by inserting adapter modules (Rebuffi et al., 2017; Houlsby et al., 2019) into each encoder and decoder layer (after the feed-forward block). Each adapter adds only a small number of new parameters, which are randomly initialized and tuned on our end task, while the rest of the network is frozen. A vanilla adapter takes as input hidden representations $[\mathbf{u_1}, h^1_1, h^1_2, \ldots, \mathbf{u_N}, \ldots, h^N_{M_N}]$, where $h^1_1, h^1_2, \ldots, h^N_{M_N}$ are textual token-level hidden representations and $\mathbf{u_1}, \ldots, \mathbf{u_N}$ are multimodal utterance-level hidden representations (in accordance to the input format presented in Figure 1a), and performs the following transformations:

$$h_{down,i} = f(\text{LN}(W_d h_i + b_d)) \tag{4}$$
$$h_{up,i} = W_u h_{down,i} + b_u \quad h_i = h_i + h_{up,i} \tag{5}$$

where $W_d \in \mathbb{R}^{d_m \text{x} d_\mathcal{B}}$, d_m is the model dimension, $d_\mathcal{B}$ is the bottleneck dimension of the adapter, $f(\cdot)$ is a non-linearity, LN a trainable layer normalization, $W_u \in \mathbb{R}^{d_\mathcal{B} \text{x} d_m}$, b_d, and b_u are the corresponding bias vectors, and $h_{down,i}$ and $h_{up,i}$ are down and up projections of h_i.

In this work, we augment the vanilla adapters of the *encoder* with a hierarchical structure (illustrated in Figure 1b). After computing (low level) self-attention between all input *textual tokens* in an encoder layer, we add a hierarchical adapter to compute *higher-level interactions* between *utterance-level multimodal* representations. By including this interaction block in the adapter, we can better propagate long-range dependencies between utterances and enforce a more global view of the events in an episode and their associations, while keeping the number of trainable parameters low.

Using the scaled dot product, we compute interaction (aka similarity) matrix H between utter-

ances (see Figure 1b) based on their *multimodal representations* $[m_1, m_2, \ldots, m_N]$:

$$e_{ij} = (W_i m_i + b_i)(W_j m_j + b_j)/\sqrt{d_m} \quad (6)$$

where W_i, W_j are learnable projection matrices, d_m is the model dimension, and e_{ij} is the degree of similarity between m_i and m_j.

At each adapter layer of the encoder, after down-projecting all vectors to the bottleneck dimension, we further contextualize utterance-level multimodal representations $u_{down,i}$ with respect to each other given the degree of similarity provided by H ("Contextualize" block in Figure 1b):

$$u'_{down,i} = \sum_{k=1}^{N} \mathrm{r}(H_{ik}/\tau) u_{down,k} + u_{down,i}$$

where N is the number of utterances, $\mathrm{r}(\cdot)$ is the softmax function, and τ is a low temperature parameter (< 1) for increasing sparsity. After contextualization, we up-project all vectors to the original dimension d_m, as in vanilla adapters (Equation (5)).

5 Content Selection

As explained earlier, episodes in SummScreen^{3D} are very long (∼5,720 tokens). BART, which has a maximum token length of 1,024, can approximately encode one fifth of the transcript.[5] We therefore perform content selection, i.e., identify salient utterances and give these as input to BART. We describe below three approaches inspired by information retrieval, summarization (Gehrmann et al., 2018; Liu and Lapata, 2019a), and computational narrative analysis (Papalampidi et al., 2021b,a).

Retrieval-based Selection We follow previous approaches (Zhang et al., 2021) in determining salient content with BM25 (Robertson and Zaragoza, 2009). BM25 is a widely known retrieval model similar to tf*idf. It assigns each utterance a "relevance" score (by comparing it against the entire transcript). Utterances with high scores are deemed salient and the K best ones are selected.

Learning-based Selection Alternatively, we may also model content selection as a binary classification problem. Given a transcript containing N utterances we predict whether each should be selected as input for the downstream summarization task (label 1) or not (label 0). We create noisy labels by matching transcript utterances to (reference) summary sentences. Specifically, we encode sentences and utterances via Sentence-BERT (Reimers and Gurevych, 2019), and assign a positive label to the utterances most similar to the reference sentences. A content selector is then trained on these pseudo-labels to identify salient utterances. We can also incorporate multimodal information in this content selection setting, using the same utterance-level representations fed into BART. We first contextualize them via a shallow transformer encoder, and add a classification head for predicting important utterances. The model is optimized with binary cross-entropy loss. During inference we select the top K predicted utterances.

Turning Point Identification We also perform content selection based on a Turning Point (TP) identification model (Papalampidi et al., 2021b,a) pre-trained on the TRIPOD movie dataset (Papalampidi et al., 2019). TPs are key events in narratives; they are distinguished into five different types depending on their functionality (e.g., Opportunity, Change of Plans, Point of No Return, Major Setback, Climax). The TP identification model considers the same multimodal information as the content selector above and identifies utterances that represent each TP. We consider the top $K/5$ predicted utterances per turning point.

6 Experimental Setup

Implementation Details We provide details of the multimodal feature extraction (i.e., utterance-level visual, audio, and textual features) in Appendix B.1. We corrupt the textual input and use the auxiliary eMLM loss (Section 4.3) only for the first X =1,500 training steps; we train our model for a total of 12,000 steps. During corruption, we mask out all content words (i.e., named entities, verbs, and nouns) and a random 10% of the input utterances. For generating summaries during inference, we use beam search with $\mathrm{beam} = 5$ and 3-gram blocking (Paulus et al., 2018). We provide further implementation details in Appendix B.2.

Training vs. Inference Although we experiment with different content selection methods during inference, we randomly sample input utterances during training. Random sampling acts as data augmentation, since the model sees slightly different input-output pairs during training at different iterations. We experimentally verify in Section 7 this

[5] We can extend positional embeddings to 1,536 by applying bilinear interpolation, however, the memory requirements would still be prohibitive for longer sequences.

is preferable to a fixed selection of utterances, especially considering the small size of our dataset. We select $K = 60$ utterances to feed into BART models given the input length limit, and order them according to their original position in the transcript.

Evaluation Metrics We evaluate the generated summaries using ROUGE F1 (Lin, 2004) against reference summaries.[6] Since ROUGE is not always a good indicator of summary quality and does not discriminate between different error types (e.g., factuality vs. fluency), we consider additional metrics based on Question-Answering (QA).[7] We obtain questions based on gold summaries and evaluate whether the correct answers exist in the generated summaries. We expect factual summaries to answer a high percentage of questions.

As in previous work (Maynez et al., 2020; Kryscinski et al., 2020; Honovich et al., 2021), we automatically generate QA pairs against reference summaries. We identify named entities and nouns using spaCy (Honnibal and Montani, 2017), and feed them as gold answers alongside the summaries to a question generator. We discriminate between named entities and nouns as answer types for measuring factuality in event-entity associations and other attributes pertaining to nouns. We used T5-base (Raffel et al., 2020) as our question generator and RoBERTa-base (Liu et al., 2019) as the QA system for answering questions given system generated summaries as input passages. Both were fine-tuned on SQuAD2.0 (Rajpurkar et al., 2016).

We measure accuracy as the partial overlap between gold and predicted answers for named entities. For nouns, we resort to textual entailment in order to account for synonyms and paraphrases in the generated summaries. We concatenate the question with gold or generated answer and predict a score for the directional relation between them. If the score is above 0.5, we consider the generated answer correct. We used BART-large (Lewis et al., 2020) fine-tuned on the MultiNLI corpus (Williams et al., 2018) as our entailment model.

We created a test suite of gold QA pairs, by retaining only those that can be answered correctly by the QA model given the reference summaries (Honovich et al., 2021). We overall generated 2,513 questions for named entities and 381 questions for nouns for the 296 episodes in our test set. On average, we have 8.5 questions per episode for named entities and 2.3 questions for nouns.[8]

[6] https://pypi.org/project/py-rouge/

[7] We also experimented with BERTScore (Zhang et al., 2020b) but observed no discernible performance differences between any pair of models.

Selection	R-2		R-L	
	text	+H-3D	text	+H-3D
Lead	6.51	—	30.72	—
Last	6.41	—	30.59	—
Middle	6.70	—	31.03	—
Random	6.54	7.24	30.91	32.15
Retrieval	6.30	6.89	30.20	31.42
TP identification	6.78	7.36	31.24	32.01
Learned selection	6.74	**7.62**	31.22	**32.64**
Pseudo-oracle	*7.96*	*8.42*	*32.85*	*33.40*

Table 3: Content selection methods for text-only BART and our multimodal Hierarchical3D variant (H-3D).

7 Results

Content Selection Table 3 compares how different approaches to content selection influence summarization performance according to ROUGE F1. We compare some simple baselines like selecting the Lead, Middle, and Last 60 utterances from the transcript as well as at Random. In addition, we compare a text only summarizer against our Hierarchical3D model. Differences amongst content selection methods are generally small. BM25 performs worse than random whilst a multimodal content selector trained on pseudo-labels performs overall best. As an upper bound, we also report results with oracle labels as input demonstrating that there is still room for improvement.

Regardless of how content is selected, we observe that our Hierarchical3D variant significantly improves performance, and interestingly, the performance gap is larger when the selection method is weaker (e.g., random vs. pseudo-oracle). This indicates that to a certain extent multimodal information makes up for suboptimal content selection.

Text vs. Multiple Modalities In Table 4 we compare our multimodal model (with the best performing content selector) against textual summarizers developed for processing long input or specifically for dialogue summarization. These include Longformer (LED; Beltagy et al. 2020) with full fine-tuning[10], a variant of LED pre-trained on

[8] We release our test suite of gold QA pairs together with the SummScreen^{3D} corpus.

[9] Textual summarizers are initialized with the same checkpoint, while some models are further tuned (e.g.,DialogLED).

[10] Adding (and tuning) adapter layers in LED led to significantly inferior performance, which in turn suggests that adapting such a network is not straightforward.

Models	R-1	R-2	R-L
HERO FT	21.56	1.74	21.27
Summ^N FT	24.71	4.42	22.61
LED FT	33.53	**7.60**	31.77
DialogLED FT	32.66	7.38	31.12
BART FT	32.61	6.94	30.83
BART AT	33.27	6.74	31.22
BART AT + H-3D	**34.51**	**7.62**	**32.64**

Table 4: Comparison of our model (BART AT + H-3D) with a video captioning model (i.e., HERO) and text-only summarizers for long dialogue summarization[9]. For HERO and all BART variants we perform content selection (FT: full fine-tuning, AT: adapter-tuning).

Models	Acc (NEs)		Acc (NNs)	
	text	+H-3D	text	+H-3D
LED FT	20.89	—	37.95	—
DialogLED FT	21.09	—	36.22	—
Summ^N FT	18.03	—	34.91	—
Random	20.25	23.64	33.86	38.06
TP identification	21.65	24.07	40.42	**40.68**
Learned selection	20.65	**24.71**	38.58	39.37
Pseudo-oracle	*28.53*	*29.64*	*41.73*	*42.00*

Table 5: QA evaluation (test set) on named entities (NEs) and nouns (NNs). We denote our Hierarchical3D model with H-3D.

dialogues (DialogLED; Zhong et al. 2022), and Summ^N (Zhang et al., 2022), a two-stage hierarchical approach for long dialogue summarization. We also present text-only BART variants, with full fine-tuning (FT) and adapter-tuning (AT). Finally, we include a SOTA video-to-text model (HERO; Li et al. 2020) with a massively pre-trained encoder, which is tuned on another TV dataset for video captioning of short clips (i.e., TVC; Lei et al. 2020b).

As can be seen in the second block of Table 4, tuning only the adapter layers (BART AT) does not hurt performance compared to full fine-tuning (BART FT), presumably due to the small dataset size. Addition of multimodal information with hierarchical adapters (BART AT + Hierarchical3D) yields substantial ROUGE improvements. Interestingly, our performance is superior to fully fine-tuned, memory-heavy models like LED or DialogLED that process the entire transcript as input. This suggests that representations from multiple modalities are more informative and lead to higher performance compared to efficient self-attention mechanisms. Summ^N performs demonstrably worse than one-stage methods and HERO fails to produce long fluent outputs due to the shallow under-trained decoder and small dataset size.

Modality	R-1	R-2	R-L
Text	34.74	7.11	32.46
Audio	33.95	6.92	31.90
Video	34.86	7.24	32.73
Multimodal	**34.95**	**7.51**	**33.01**
w/ vanilla adapters	34.25	7.45	32.41
w/o eMLM loss	33.80	6.84	31.88
w/o random augmentation	33.45	6.48	31.81

Table 6: The role of multimodal information and hierarchical adapters (validation set).

QA Evaluation The results of our automatic QA evaluation are summarized in Table 5. The second block focuses on model performance with different content selection variants. We only compare text-only and multimodal (+H-3D) BART. Again, we find that augmenting BART with multimodal information regardless of the selection method improves accuracy, especially for named entities. This is true even when content is selected by a pseudo-oracle suggesting that multimodal information provides better associations between events and entities, even when the input contains all salient information. We further observe that supervised content selection and TP identification offer the best performance. The first block reports the performance of state-of-the-art models on dialogue summarization; we find these models perform on par or slightly worse than textual BART (depending on the content selection method) which casts doubts on their ability to efficiently consume longer inputs. Examples of output summaries (and QA pairs) are given in Table 7 and Appendix C.3. We also report additional (entity-specific) results in Appendix C.2.

Ablation Studies In Table 6 we summarize our ablation studies which isolate the contribution of individual modeling components. We observe that individual modalities (Text, Audio, Video) are worse on their own than in combination (Multimodal). The least informative modality is audio, while the most informative is video. In the multimodal case, we substitute hierarchical adapters in the encoder with vanilla adapters and observe a small drop in performance. Removing the auxiliary eMLM loss during training further decreases performance. The auxiliary loss is crucial for rendering the textual encoder multimodal and forcing an already tuned summarizer to consider a different type of input. Finally, data augmentation (via random content selection) during training is also important given the small size of our dataset and BART encoder length restrictions. We report additional ablation

Model	Summary	
Gold	Joshua tells Elizabeth he wants to turn Allison and demands she help ease Allison into her new life as his wife. Elizabeth tells Joshua she will kill him before she allows him to hurt Allison. Livvie is able to fend off her need to feed while she and Caleb make love. Frank searches for Allison. When Frank attempts to kidnap Allison from Rafe, he discovers that it really is Lucy and I an in disguise. Allison and Rafe reappear in Caleb s cave.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Rafe**
	• Who does Frank try to kidnap?	**Allison**
	• Who tries to kidnap Allison?	**Frank**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**wife**
Text-only	Rafe tells Alison that he will never let Joshua take her for his bride, but she tells him that she has no choice in the matter. Elizabeth tells Joshua that she will not stand by and allow him to take her daughter. Joshua tells Elizabeth that he is going to eas e Alison into her new lifestyle as his wife. Elizabeth says that she is not going to let her daughter suffer the kind of nightmare that she lived. She will kill Joshua before he is even that close to turning her. Alison tells Rafe that she thinks this is a little extreme, that is all. Rafe says he will not let Joshua get to her. He promises to keep her away from Joshua and all his goons. Caleb tells Livvie that she doesn t need to feed. He tells her that he can t make love to her the way she wants to. She tells him she can t turn him back, but he tells her he can. He says that he loves her and that he wants to make her his bride.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Joshua**
	• Who does Frank try to kidnap?	**Joshua**
	• Who tries to kidnap Allison?	**Rafe**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**wife**
H-3D	Livvie tries to convince Caleb to let her take the risk of biting him, but she is afraid that she won t be able to do it. Joshua tells Elizabeth that he wants Alison to be his bride. Elizabeth is shocked when she finds out that Joshua wants to take Alison away from Rafe. Elizabeth tells Joshua that she will find a way to stop him from taking Alison. Rafe tells Alison that he has a vision of a city of tortured souls, a master vampire race, and his vampire bride. He tells her that he can make a perfect bride for her. Alison tells Rafe that she doesn t want to leave her family, but Rafe assures her that she is not going to leave them. Frank tells Ian that he is going to have to tell his boss that his mission didn t work.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Rafe**
	• Who does Frank try to kidnap?	**Allison**
	• Who tries to kidnap Allison?	**Rafe**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**vampire bride**

Table 7: Examples of gold and model generated summaries together with automatically generated questions and their answers based on gold and automatic summaries. Correct/wrong answers are in green/red color. We show output for a text-only BART model and a multimodal variant with hierarchical adapters (H-3D); in both cases content selection is performed with a model trained on pseudo-labels.

experiments on content selection in Appendix C.1.

8 Conclusions

In this work, we addressed the task of multimodal abstractive summarization and created SummScreen^{3D}, a new dataset which we hope will facilitate future research in this direction. We incorporated multimodal information into a pre-trained textual summarizer in a parameter-efficient manner and have experimentally shown performance gains over text-only models. Our experimental results further underscore the importance of (multimodal) content selection compared to approaches focusing on self-attention variants for long dialogue summarization. In the future, we plan to explore more *structure-aware* representations for *all* input modalities in order to improve factuality (e.g., entity-event associations) in the generated summaries.

Acknowledgements

We thank the anonymous reviewers for their feedback. The authors also thank Marcus Rohrbach and Frank Keller for insightful comments on earlier versions of this paper. We gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (grant EP/W002876/1).

9 Limitations

Our approach considers only coarse-grained (i.e., utterance-level) multimodal information which we demonstrate is beneficial for summarization. More detailed frame-level visual information e.g., person identification and object recognition in frames, would be useful. However, considering frame-level representations for hour-long videos would bring a considerable increase in memory requirements and additional difficulties in aligning different modalities (e.g., frames vs. tokens vs. audio segments). We leave these challenges to future work and believe that structure-aware methods are necessary for addressing the current limitations.

Following previous work (Maynez et al., 2020; Kryscinski et al., 2020; Honovich et al., 2021), we advocate the use of automatic QA-based methods for evaluating the generated summaries. Although there is supportive analysis (e.g., Tang et al. 2022) that shows better correlation to human judgements for QA-based automatic evaluation in comparison with traditional summarization metrics such as ROUGE, more experimentation is necessary to determine the shortcomings of these metrics.

Finally, conducting human evaluation for SummScreen^{3D} is infeasible, since this would entail asking judges to watch 40-minute long episodes in order to evaluate the content and faithfulness of the summaries. We further cannot assume judges are familiar with the characters, specific details and (complex) storylines of different soap operas contained in our test set in order to be able to make reliable judgments. Therefore, using QA-based metrics for judging specific attributes of summarization quality, such as whether the correct entities are linked to the correct events in an episode (i.e., QA evaluation related to named entities), can provide us with useful insights.

References

Christos Baziotis, Ivan Titov, Alexandra Birch, and Barry Haddow. 2021. Exploring unsupervised pretraining objectives for machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2956–2971, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc.

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)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 6201–6210, Seoul, Korea (South). IEEE Computer Society.

Cheng-Yang Fu, Joon Lee, Mohit Bansal, and Alexander Berg. 2017. Video highlight prediction using audience chat reactions. In *Proceedings of the 2s017 Conference on Empirical Methods in Natural Language Processing*, pages 972–978, Copenhagen, Denmark. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780. IEEE Computer Society.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-

annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Proceedings of the 13th European Conference on Computer Vision*, volume 8695 of *Lecture Notes in Computer Science*, pages 505–520, Zurich, Switzerland. Springer.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *Sentometrics Research*, 7(1):411–420.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799, Long Beach, California, USA. PMLR.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR*, San Diego, CA, USA.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *CoRR*, abs/2001.04451.

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)*, pages 9332–9346, Online. Association for Computational Linguistics.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020a. MART: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2603–2614, Online. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020b. TVR: A large-scale dataset for video-subtitle moment retrieval. In *Proceedings of the 16th European Conference on Computer Vision*, pages 447–463. Springer.

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*, pages 7871–7880, Online. Association for Computational Linguistics.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.

Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Wang, William Yang Wang, Tamara L. Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, and Zicheng Liu. 2021. VALUE: A multi-task benchmark for video-and-language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. 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.

Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019b. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *CoRR*, abs/2002.06353.

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

Cory S Myers and Lawrence R Rabiner. 1981. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *Bell System Technical Journal*, 60(7):1389–1409.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada. Association for Computational Linguistics.

Shashi Narayan, Shay Cohen, and Maria Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807. Association for Computational Linguistics.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy. Association for Computational Linguistics.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1707–1717, Hong Kong, China. Association for Computational Linguistics.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2021a. Film trailer generation via task decomposition. *CoRR*, abs/2111.08774.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2021b. Movie summarization via sparse graph construction. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 13631–13639. AAAI Press.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the 6th International Conference on Learning Representations, ICLR*, Vancouver, BC, Canada. OpenReview.net.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Processings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. volume abs/1705.08045.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *Proceedings of the 36th German Conference on Pattern Recognition*, volume 8753 of *Lecture Notes in Computer Science*, pages 184–195, Münster, Germany. Springer.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123(1):94–120.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In *In Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL), NIPS.*

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: standardized comparison over long language sequences. *CoRR*, abs/2201.03533.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. TVSum: Summarizing web videos using titles. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187, Boston, MA, USA. IEEE Computer Society.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. VL-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, New Orleans, LA, USA.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, Las Vegas, NV, USA.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. 2020. Sparse sinkhorn attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9438–9447, Virtual Event. PMLR.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems 34*, pages 200–212. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021. VLM: task-agnostic video-language model pre-training for video understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4227–4239, Online Event. Association for Computational Linguistics.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, Las Vegas, NV, USA.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339. PMLR.

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

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4426–4433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. DialogLM: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 11765–11773, Virtual Event. AAAI Press.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 7590–7598, New Orleans, Louisiana, USA. AAAI Press.

As The World Turns (atwt)	1356
Bold and the Beautiful (bb)	1113
Guiding Light (gl)	836
One Life to Live (oltl)	1118
Port Charles (pc)	501

Table 8: Distribution of different TV shows in the augmented dataset.

	TMS	SummScreen^{3D}
TV shows	10	5
Episodes	22,503	4,575
min #episodes per show	168	501
max #episodes per show	3,784	1,356
median #episodes per show	1,973.5	1,113
avg #episodes per show	2,250.0	984.8
Utterances/episode	360.8	322.8
Tokens/episode	6,420.7	5,720.6
Summaries/episode	1	1.53
#tokens/summary	327.0	395.7

Table 9: Comparison between the original SummScreen-TMS (Chen et al., 2022)) and SummScreen^{3D} which is a subset of the original dataset.

A Dataset Analysis

As mentioned in Section 3, we create a multimodal version of the SummScreen dataset (Chen et al., 2022) by collecting full-length videos of the episodes contained in the original dataset. Overall, we retrieved videos from YouTube for five different TV shows (i.e., soap operas). We present in Table 8 the names of the TV shows and the number of episodes per show. We made sure to have enough episodes from each TV show and maintain the same distribution when splitting the dataset into train, validation, and test. Moreover, we present an example of the input transcript and output summary from SummScreen (Chen et al., 2022) and how we augment the dataset with additional information from the full-length video in Figures 2 and 3.

Next, we also compare the statistics of SummScreen^{3D}, which is a subset of SummScreen-TMS (Chen et al., 2022), with the original dataset in Table 9. Overall, we include episodes from half the TV shows contained in TMS. The number of episodes per TV show in our dataset is more balanced in comparison with the original (see rows 3–6 in Table 9). SummScreen^{3D} has similar input and output statistics per episode to the original dataset (e.g., number of utterances and tokens per transcript and number of tokens per summary). However, we also collect more summaries per episode when available (see Table 8) for creating an augmented training set and a more robust evaluation set.

Finally, we also compare our dataset against other video-to-text summarization datasets in Table 10. TACoS (Rohrbach et al., 2014) and How2 (Sanabria et al., 2018) are the only available multimodal summarization datasets we are aware of. In comparison, our dataset contains much longer videos (on average 40 minutes long), and fairly elaborate textual summaries (instead of short one-sentence descriptions with simple vocabulary).

B Implementation Details

B.1 Dataset Pre-processing

Given full-length video, we extract features for all modalities at the utterance-level as mentioned in Section 4.2. For text, we extract sentence-level features using Sentence-BERT (Reimers and Gurevych, 2019). Each utterance in the transcript is thus represented by a fixed-size vector. For the frames, we extract two types of features: frame-level features using the CLIP visual encoder (Radford et al., 2021) and motion-level features from video clips using Slowfast (Feichtenhofer et al., 2019). We then aggregate frame- and motion-level features to utterance-level given the automatic alignment by mean pooling. Finally, for audio, we use YAMNet pre-trained on the AudioSet-YouTube corpus (Gemmeke et al., 2017) for classifying audio segments into 521 audio classes (e.g., tools, music, explosion); for each audio segment contained in a shot, we extract features from the penultimate layer, and then aggregate representations again to utterance-level via mean pooling.

B.2 Training Details

We used the Adam algorithm (Kingma and Ba, 2015) for optimizing our networks. We trained all models with a learning rate of $3e{-}5$ for 12k steps using a linear warm-up of 500 steps, followed by inverted squared decay. All BART-based models were trained with batch size of 1 episode on 4 P100 GPUs with 16GB memory and label smoothing (Szegedy et al., 2016) of 0.1. To fine-tune the LED-based models, we used 4 A100 GPUs with 80GB memory. It took approximately 12 hours to fully train each of these models. Fully fine-tuned models have 406M parameters, which are all fine-tuned on the target dataset, whereas our multimodal adapter-augmented model has 421.6M parameters, from which we only train 15.6M parameters (i.e., multimodal projection layer and adapter layers) on the target dataset. This means that we only tune $\sim$3.8% of the model parameters of the fully fine-tuned models. We report the results of a single run for all models following previous work (Chen et al., 2022; Zhong et al., 2022) due to the computational overhead of running some large comparison models. However, we report in Table 11 the average and standard deviation over three runs for BART AT and BART AT + H-3D in order to demonstrate the performance variation of these models.

C Additional Experimental Results

C.1 Ablation Study on Content Selection

In Table 12, we examine the performance of different content selectors. We report precision (Pre), recall (Re), and F1 score of model variants based on pseudo-oracle labels. We first consider selectors which have not been trained with pseudo-oracle labels, such as Random, Retrieval (i.e., BM25) and TP identification (we refer to these approaches as unsupervised). We observe that unsupervised baselines have significantly lower F1 score in comparison with a supervised approach. Interestingly, although TP identification "agrees less" with the pseudo-oracle labels in comparison with BM25, TPs still present competitive performance against the supervised content selector on abstractive textual summarization (e.g., Table 5). Finally, comparing the multimodal supervised content selector with equivalent unimodal models, we observe that the highest performance is achieved by combining all modalities. With respect to unimodal variants, we find that the textual modality is most informative, while using visual or audio cues alone is not enough to predict salient content.

C.2 Entity-specific Evaluation

Chen et al. (2022) propose a set of entity-specific metrics in order to investigate the role of characters, which are fundamental in TV shows, in the generated summaries. Specifically, they measure several bag of character (BoC) metrics based on character overalp between generated and gold standard summaries. They define precision as the fraction of correctly mentioned characters with respect to all characters that appear in the generated summary (BoC-p) and recall as the fraction of correctly mentioned characters with respect to all characters that appear in the gold summary (BoC-r). Given precision and recall, we also measure F1-score (BoC-f1).

Apart from correctly mentioned characters, Chen

	dataset size	video input	text input	video duration	output tokens
TACoS	147	✓	✗	4.5 minutes	9
How2	79k	✓	✓	90 seconds	20
SummScreen^{3D}	4.5k	✓	✓	40 minutes	290

Table 10: Comparison between SummScreen^{3D} and other video-to-text summarization datasets (see Table 1).

	R-2	R-L
BART AT	6.71 (0.02)	30.96 (0.23)
BART AT + H-3D	7.58 (0.03)	7.58 (0.03)

Table 11: Results of two models from Table 3 across three different runs. We report the average and standard deviation in parentheses for R-2 and R-L.

Unsupervised	Precision (%)	Recall (%)	F1 (%)
Random	19.55	20.90	20.06
Retrieval	24.63	26.62	25.40
TP identification	20.35	22.10	21.04

Supervised	Precision (%)	Recall (%)	F1 (%)
Multimodal	**47.57**	**50.68**	**48.57**
Text	45.26	48.54	46.52
Vision	22.97	24.91	23.73
Audio	21.54	23.29	22.23

Table 12: The role of multimodal information in content selection. We report the Precision, Recall, and F1 for selecting important utterances from an episode. Supervised models are trained on pseudo-oracle labels.

et al. (2022) also compute similar bag of words metrics for relations between characters in the summaries. Specifically, they consider a pair of characters related if they appear in the same sentence in the summary. They do not account for the direction of relations and focus only on co-occurrence. They again consider precision (BoR-p) and recall (BoR-r) of the intersection of pairs of characters similarly to computing the BoC metrics. We also report F1-score (BoR-f1), given the precision and recall for character relations.

We summarize our entity-specific results in Table 13. Overall, especially when considerring the F1 scores for characters and relations, we arrive to similar conclusions as with our automatic QA evaluation (Table 5). The multimodal information that is incorporated in our Hierarchical3D approach increases most entity-specific metrics in comparison with text-only variants. Regarding different content selection methods, TP identification and supervised content selection again perform best in comparison with random selection, although differences are not large. Finally, we achieve the best F1 scores in both entity- and relation-specific metrics by using oracle selection, indicating that there is still room for improvement. Interestingly, we again observe a further increase in performance by adding multimodal information in the pseudo-oracle variant, suggesting that video-based information is important even when we consider the most salient parts of an episode.

We also compare our approach with state-of-the-art, fully fine-tuned textual summarizers for long dialogues. We again notice that Summ^N is weakest according to entity-specific metrics. Next, efficient architectures for modeling the entire input (i.e., LED, DialogLED) have competitive performance against our text-only variants with content selection. However, Hierarchical3D that considers multimodal information outperforms these memory-heavy models while training only a small fraction of model parameters. This further validates our hypothesis that the video can provide additional information which more important for high-quality summaries than processing the entire textual input.

C.3 Examples of Generated Summaries

In this section we provide examples of generated summaries based on different automatic systems. Moreover, we provide examples of questions and answers used for the automatic QA evaluation described in Section 6.

Table 14 shows examples of automatically generated question-answer pairs given gold standard summaries. We provide examples of QA pairs for named entities (first 4 rows of the table) and nouns (remaining 6 rows of the table). We observe that most QA pairs are reasonable and correspond to information given in human-written summaries (first column of the table). However, there are cases where the QA pairs do not provide reasonable questions. Such an example is illustrated in the last row of Table 14, where the question is generated given the summary segment "Jonathan and Lizzie find out their baby has a medical condition, and make a run for it":

Q: "What do Lizzie and Jonathan do when they learn their baby has a medical condition?"

A: "run"

	BoC-p	BoC-r	BoC-f1	BoR-p	BoR-r	BoR-f1
Random selection	82.55	38.71	52.71	29.82	9.39	14.28
+ Hierarchical3D	81.80	47.37	60.00	31.75	13.77	19.21
TP identification	**84.31**	38.93	53.26	**36.79**	10.33	16.13
+ Hierarchical3D	82.20	47.10	59.89	34.82	**14.10**	**20.07**
Content Selection	81.60	36.59	50.52	30.54	8.58	13.40
+ Hierarchical3D	81.90	**48.48**	**60.91**	33.04	**14.37**	**20.03**
Pseudo-oracle	*87.42*	*46.95*	*61.09*	*37.92*	*14.40*	*20.87*
+ Hierarchical3D	*85.53*	*52.37*	*64.96*	*36.67*	*17.51*	*23.70*
LED FT	82.28	33.54	47.65	34.35	10.64	16.25
DialogLED FT	82.93	38.19	52.27	31.71	10.32	15.57
Summ^N FT	82.74	29.14	43.10	34.73	9.39	14.78

Table 13: Entity-specific metrics (test set). We report bag of character precision (BoC-p), recall (BoC-r), and F1 (BoC-f1). Analogously, we compute bag of relations precision (BoR-p), recall (BoR-r), and F1 (BoR-f1).

Summary	Question	Answer
Sage goes to live with Jack after she learns Carly is planning to marry Craig. Meg agrees to marry Dusty.	Who does Meg agree to marry?	Dusty
	Who does Sage go to live with?	Jack
Joshua is busy preparing for Allison's arrival, as he unveils Kevin's latest creation; a portrait of Allison and Joshua in their wedding attire. Lucy goes to church to plead for answers. Ian overhears her plea and swears that he will not let her die. Livvie shows Joshua a picture of Allison appearing to be dead and tells him that he was right her fangs are poisoned.	Who goes to church to plead for answers?	Lucy
	Who swears he will not let Lucy die?	Ian
	What does Lucy do at church?	plea
	What part of Allison's body is poisoned?	flangs
Lizzie and Jonathan spend some time with their baby. Jonathan gives in to one of Alan s demands. Gus and Harley find a disk with some interesting information on it. Gus still can t figure out what it is that Blake has on him. Dinah and Mallet argue over who will be the next WSPR star. Tammy is heartbroken after a visit to the hospital. Jonathan and Lizzie find out their baby has a medical condition, and make a run for it. Alan realizes that he may have been outwitted by Jonathan. Gus vows to get to the bottom of his supposed secret.	What does Gus vow to find out about Blake?	secret
	What is Lizzie and Jonathan spending time with?	baby
	What do Gus and Harley find?	disk
	What do Lizzie and Jonathan do when they learn their baby has a medical condition?	run

Table 14: Examples of automatically generated QA pairs for the evaluation of generated summaries.

This QA pair does not correspond to a reasonable fact of the episode. This shows that although it is useful to filter the questions, there are still imperfections with the automatic generation of QA pairs, especially when considering nouns.

Next, we give examples of the generated summaries for the TV show "Port Charles" in Tables 15–18. We present the gold or generated summary alongside the QA pairs used for evaluation. First, we compare different content selection methods (i.e., supervised content selection (CS), TP identification (TPs), and pseudo-oracle) for a text-only summarizer based on BART with adapter tuning. We present two examples in Tables 15 and 17 (we also show gold summaries for each episode). In both cases, we observe that the pseudo-oracle selection provides summaries of better quality, with fewer errors in the questions answered (i.e., errors are illustrated with red). Moreover, when comparing content selection (CS) with TP identification (TPs), we find that these two approaches provide similar results, as suggested by our main experimental results (Table 5). Specifically, in Table 15, TP identification seems to provide the most informative summary, whereas in Table 17 supervised content selection is the best option.

Secondly, we compare our approach that considers multimodal information (Hierarchical3D) against text-only BART with equivalent content selection, and LED which considers only text and uses an efficient self-attention mechanism for processing the entire input. We present two examples for the same episodes as above in Tables 16 and 18. We empirically validate that the quality of the generated summaries is improved by adding the multimodal information (both when using supervised content selection and TP identification). Our approach leads to summaries that answer a

larger percentage of automatic questions correctly (i.e., correct answers are illustrated with green) outperforming LED, which is fully fine-tuned and memory-heavy. Interestingly, LED summaries cannot answer a large proportion of the given questions, suggesting that such methods may not be suitable for our task and small size dataset.

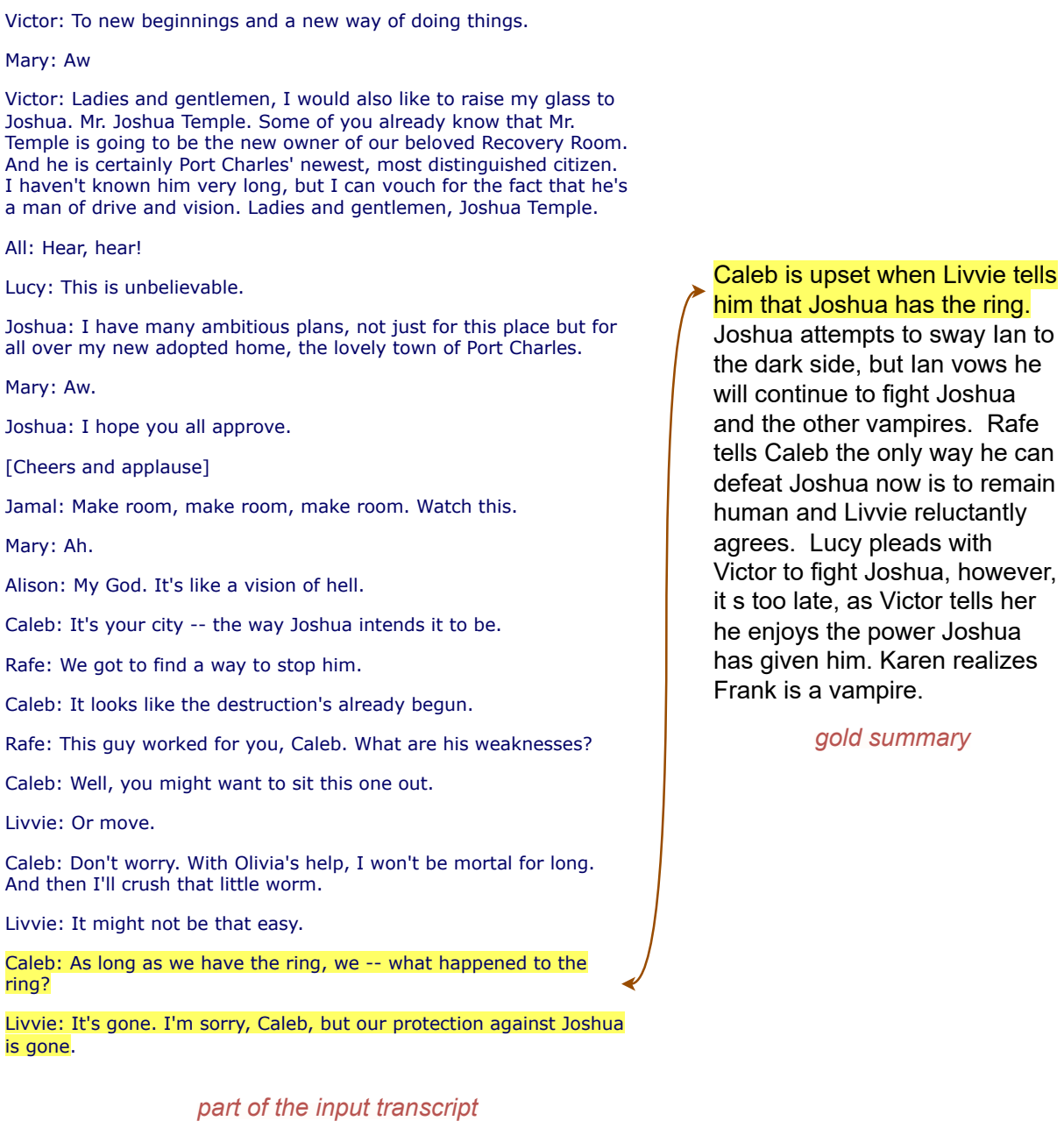


Figure 2: Example of input and output for SummScreen dataset. A long transcript is considered as input for summarization, containing the dialogue parts of a full-length TV episode. Character names are given as part of the dialogue. The goal is to produce a textual summary of most important events in the episode.

Victor: To new beginnings and a new way of doing things.

Mary: Aw

Victor: Ladies and gentlemen, I would also like to raise my glass to Joshua. Mr. Joshua Temple. Some of you already know that Mr. Temple is going to be the new owner of our beloved Recovery Room. And he is certainly Port Charles' newest, most distinguished citizen. I haven't known him very long, but I can vouch for the fact that he's a man of drive and vision. Ladies and gentlemen, Joshua Temple.

All: Hear, hear!

Lucy: This is unbelievable.

Joshua: I have many ambitious plans, not just for this place but for all over my new adopted home, the lovely town of Port Charles.

Mary: Aw.

Joshua: I hope you all approve.

[Cheers and applause]

Jamal: Make room, make room, make room. Watch this.

Mary: Ah.

Alison: My God. It's like a vision of hell.

Caleb: It's your city -- the way Joshua intends it to be.

Rafe: We got to find a way to stop him.

Caleb: It looks like the destruction's already begun.

Rafe: This guy worked for you, Caleb. What are his weaknesses?

Caleb: Well, you might want to sit this one out.

Livvie: Or move.

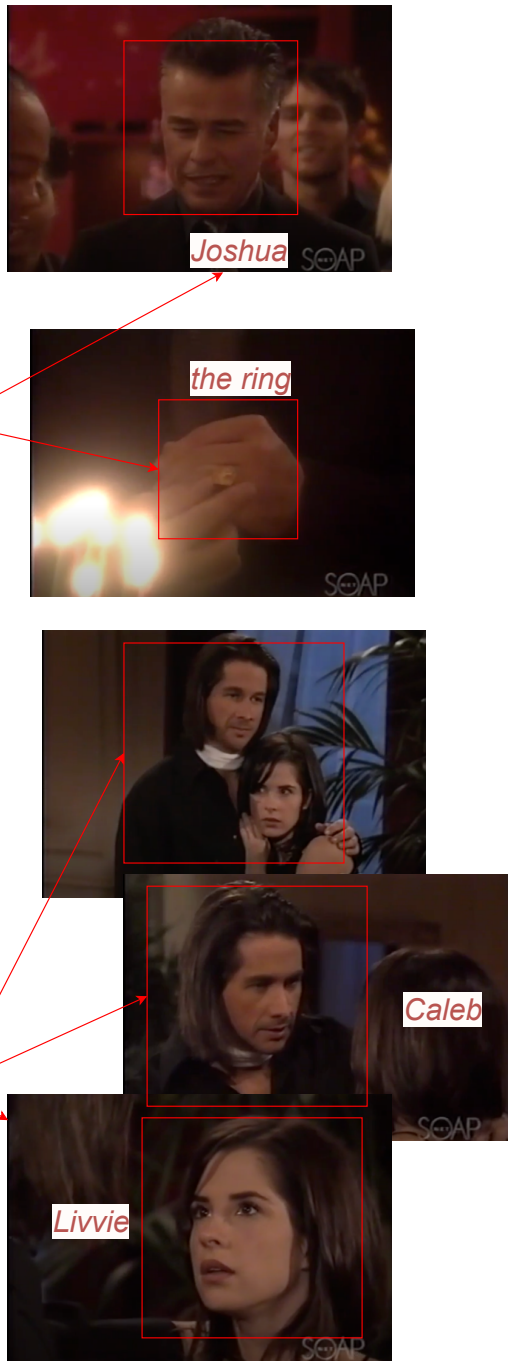
Caleb: Don't worry. With Olivia's help, I won't be mortal for long. And then I'll crush that little worm.

Livvie: It might not be that easy.

Caleb: As long as we have the ring, we -- what happened to the ring?

Livvie: It's gone. I'm sorry, Caleb, but our protection against Joshua is gone.

part of the input transcript



video

Figure 3: We augment SummScreen (see example of Figure 2) with information from the full-length video, which is aligned to the input transcript. Additional information, such as Joshua touching the ring in a previous scene or Caleb looking concerned when talking to Livvie, can be acquired from the video frames.

Model	Summary	
Gold	Caleb is upset when Livvie tells him that Joshua has the ring. Joshua attempts to sway Ian to the dark side, but Ian vows he will continue to fight Joshua and the other vampires. Rafe tells Caleb the only way he can defeat Joshua now is to remain human and Livvie reluctantly agrees. Lucy pleads with Victor to fight Joshua, however, it s too late, as Victor tells her he enjoys the power Joshua has given him. Karen realizes Frank is a vampire.	
QA pairs	• Who tells Lucy that he enjoys the power Joshua has given him?	**Victor**
	• Who does Karen realize is a vampire?	**Frank**
	• Who pleads with Victor to fight Joshua?	**Lucy**
	• Who tells Caleb that Joshua has the ring?	**Livvie**
	• Who realizes Frank is a vampire?	**Karen**
	• What does Livvie tell Caleb Joshua has?	**the ring**
	• Who does Karen realize Frank is?	**vampire**
CS (text-only)	Caleb and Rafe discuss how to get close to Joshua and Livvie. Lucy tries to convince Victor that Joshua is an evil vampire who should not be allowed to have his soul. Lucy tells Victor that she can t lose him and wants him to accept her offer to turn him back into a vampire. Joshua tells the people of Port Charles that he will do whatever it takes to breathe new life into this wonderful old place.	
QA pairs	• Who tells Lucy that he enjoys the power Joshua has given him?	**Victor**
	• Who does Karen realize is a vampire?	**Joshua**
	• Who pleads with Victor to fight Joshua?	**Lucy**
	• Who tells Caleb that Joshua has the ring?	**Rafe**
	• Who realizes Frank is a vampire?	**Victor**
	• What does Livvie tell Caleb Joshua has?	**soul**
	• Who does Karen realize Frank is?	**vampire**
TPs (text-only)	Caleb and Livvie are shocked to find out that the ring has been taken away from them by Joshua. They are unable to get the ring back, but they are determined to find a way to get it back. Lucy tells Victor that Joshua is a liar and that he should not be allowed to have an important position in Port Charles. Victor tells Lucy that he will not give up on her, but she tells him that she will not go to the hospital because she has to be here for Victor. Lucy and Victor agree that they need to stay at the hospital for the sake of Victor, but Lucy is not willing to give up her job as Victor s guardian angel. Rafe tells Alison that if she turns back, she is going to be under Joshua s control, and he doesn t think she will be too happy about taking orders from someone else. He tells her that he would rather slit her own wrists than have Joshua touch her.	
QA pairs	• Who tells Lucy that he enjoys the power Joshua has given him?	**Victor**
	• Who does Karen realize is a vampire?	**Joshua**
	• Who pleads with Victor to fight Joshua?	**Lucy**
	• Who tells Caleb that Joshua has the ring?	**Lucy**
	• Who realizes Frank is a vampire?	**Lucy**
	• What does Livvie tell Caleb Joshua has?	**the ring**
	• Who does Karen realize Frank is?	**vampire**
Pseudo-oracle	Caleb and Livvie find out that the ring that protects them from Joshua is gone, and they have no way of getting it back. Lucy and Ian try to convince Victor to turn Joshua into a vampire, but he tells them that he is not a monster and that he wants to be a good man. Lucy tells Victor that Joshua is evil, but Victor tells her that she is his friend and he will not let her get to him. Victor tells Lucy that he will do whatever it takes to breathe new life into Port Charles. Ricky tells Jamal that he does not want to be his guardian angel anymore, but Jamal tells him to go and hang with the bloodsuckers. Jack tells Jack that the only way he is going to win is to play both sides of the fence. Frank tells Karen that there are benefits to his new role as a vampire and he is willing to do it for the good of the town of Port Charles and his family.	
QA pairs	• Who tells Lucy that he enjoys the power Joshua has given him?	**Victor**
	• Who does Karen realize is a vampire?	**Frank**
	• Who pleads with Victor to fight Joshua?	**Lucy**
	• Who tells Caleb that Joshua has the ring?	**Livvie**
	• Who realizes Frank is a vampire?	**Karen**
	• What does Livvie tell Caleb Joshua has?	**the ring**
	• Who does Karen realize Frank is?	**vampire**

Table 15: Examples of generated summaries and automatic QA evaluation for different content selection methods. In all model variants we only consider the textual modality (i.e., BART with adapter tuning).

Model	Summary
Gold	Caleb is upset when Livvie tells him that Joshua has the ring. Joshua attempts to sway Ian to the dark side, but Ian vows he will continue to fight Joshua and the other vampires. Rafe tells Caleb the only way he can defeat Joshua now is to remain human and Livvie reluctantly agrees. Lucy pleads with Victor to fight Joshua, however, it s too late, as Victor tells her he enjoys the power Joshua has given him. Karen realizes Frank is a vampire.
QA pairs	• Who tells Lucy that he enjoys the power Joshua has given him? **Victor** • Who does Karen realize is a vampire? **Frank** • Who pleads with Victor to fight Joshua? **Lucy** • Who tells Caleb that Joshua has the ring? **Livvie** • Who realizes Frank is a vampire? **Karen** • What does Livvie tell Caleb Joshua has? **the ring** • Who does Karen realize Frank is? **vampire**
Text-only (TPs)	Caleb and Livvie are shocked to find out that the ring has been taken away from them by Joshua. They are unable to get the ring back, but they are determined to find a way to get it back. Lucy tells Victor that Joshua is a liar and that he should not be allowed to have an important position in Port Charles. Victor tells Lucy that he will not give up on her, but she tells him that she will not go to the hospital because she has to be here for Victor. Lucy and Victor agree that they need to stay at the hospital for the sake of Victor, but Lucy is not willing to give up her job as Victor s guardian angel. Rafe tells Alison that if she turns back, she is going to be under Joshua s control, and he doesn t think she will be too happy about taking orders from someone else. He tells her that he would rather slit her own wrists than have Joshua touch her.
QA pairs	• Who tells Lucy that he enjoys the power Joshua has given him? **Victor** • Who does Karen realize is a vampire? **Joshua** • Who pleads with Victor to fight Joshua? **Lucy** • Who tells Caleb that Joshua has the ring? **Lucy** • Who realizes Frank is a vampire? **Lucy** • What does Livvie tell Caleb Joshua has? **the ring** • Who does Karen realize Frank is? **vampire**
H-3D (TPs)	Caleb and Livvie are shocked when they find out that their protection against Joshua is gone. Victor and Lucy argue about Victor's role in Port Charles. Lucy tells Victor that Joshua is evil and that he should not be allowed to have an important position with the vampires. Victor tells Lucy that he still has so much to contribute and maybe this is his chance to have people listen to him again. Lucy is upset that Victor wants to give Joshua an important role in the town. Lucy and Victor are at the hospital and Lucy tells him that she is going to be there for Victor, but he tells her to stay away from him. Frank tells Karen that he has grown a pair of fangs. Karen is shocked to hear that Frank is a vampire.
QA pairs	• Who tells Lucy that he enjoys the power Joshua has given him? **Victor** • Who does Karen realize is a vampire? **Frank** • Who pleads with Victor to fight Joshua? **Lucy** • Who tells Caleb that Joshua has the ring? **Lucy** • Who realizes Frank is a vampire? **Karen** • What does Livvie tell Caleb Joshua has? **their protection against Joshua** • Who does Karen realize Frank is? **vampire**
LED	At the end of the night, Caleb and Livvie s wedding ring is revealed to Rafe and Caleb s ring. Caleb tells Rafe that the ring is a vampire slayer. Rafe tells Caleb that he s going to be a vampire again. Lucy tells Victor that Victor is going to take over the town of Port Charles. Victor tells Lucy that he wants to talk to Lucy. Lucy asks Victor to join her in the new venture, but Victor tells her that he is not going to do it. Caleb tells Olivia that he has been drugged by Kevin, and he s been turned into a vampire. Olivia tells him that she wants to be part of the new club, but Caleb tells her to stay away from him. Joshua tells Ian that he will not be able to get Victor away from Victor. Ian tells Joshua that Joshua is not one of the vampire slayers, but he is the one of them.
QA pairs	• Who tells Lucy that he enjoys the power Joshua has given him? **Victor** • Who does Karen realize is a vampire? **Caleb** • Who pleads with Victor to fight Joshua? **Ian** • Who tells Caleb that Joshua has the ring? **Ian** • Who realizes Frank is a vampire? **Rafe** • What does Livvie tell Caleb Joshua has? **wedding ring** • Who does Karen realize Frank is? **slayer**

Table 16: Examples of generated summaries and automatic QA evaluation for different models. Here we compare our Hierarchical3D model (H-3D) with state-of-the-art textual summarizers (i.e., LED).

Model	Summary	
Gold	Joshua tells Elizabeth he wants to turn Allison and demands she help ease Allison into her new life as his wife. Elizabeth tells Joshua she will kill him before she allows him to hurt Allison. Livvie is able to fend off her need to feed while she and Caleb make love. Frank searches for Allison. When Frank attempts to kidnap Allison from Rafe, he discovers that it really is Lucy and I an in disguise. Allison and Rafe reappear in Caleb s cave.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Rafe**
	• Who does Frank try to kidnap?	**Allison**
	• Who tries to kidnap Allison?	**Frank**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**wife**
CS (text-only)	Rafe tells Alison that he will never let Joshua take her for his bride, but she tells him that she has no choice in the matter. Elizabeth tells Joshua that she will not stand by and allow him to take her daughter. Joshua tells Elizabeth that he is going to eas e Alison into her new lifestyle as his wife. Elizabeth says that she is not going to let her daughter suffer the kind of nightmare that she lived. She will kill Joshua before he is even that close to turning her. Alison tells Rafe that she thinks this is a little extreme, that is all. Rafe says he will not let Joshua get to her. He promises to keep her away from Joshua and all his goons. Caleb tells Livvie that she doesn t need to feed. He tells her that he can t make love to her the way she wants to. She tells him she can t turn him back, but he tells her he can. He says that he loves her and that he wants to make her his bride.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Joshua**
	• Who does Frank try to kidnap?	**Joshua**
	• Who tries to kidnap Allison?	**Rafe**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**wife**
TPs (text-only)	Livvie tells Caleb that she can t be with him, knowing what his bite might do to him. Joshua tells Elizabeth that he is going to steal Alison s slayer s intended and that will make him more respectable. He tells her that she is not going to do it, and that he knows that both roads lead Alison right back to Joshua. Elizabeth tells Joshua that she has no idea how much she would love to do that, but they both know that they are too busy figuring out which road to take. Rafe and Alison find themselves in Caleb s old hangout. Alison is shocked to see that Rafe is still there. She tells Rafe that she will feel better when all this is over with. She asks Rafe if he has his back and he says that he does, and he tells her he has her back. Lucy and Ian find out that they have succeeded in their mission to kidnap Alison. Ian and Frank tell Lucy that they will have to tell their boss that their mission didn t work. Frank tells Ian that he will pay for this, and they are both going to pay for it. Ian tells Frank that he has to tell his big, powerful boss that his mission failed.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Lucy**
	• Who does Frank try to kidnap?	**Alison**
	• Who tries to kidnap Allison?	**Lucy**
	• Who can fend off her need to feed while she and Caleb make love?	**Rafe**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**Caleb**
Pseudo-oracle	Caleb tells Livvie that he loves her and wants to make love to her, but he can t because his fangs are poisoned and he can s not turn her back. He tells her to take the risk, but she is afraid to do it. She tells him that she loves him and that she wants to be with him, but that she can t do it because of the risk of his bite. He says that he will not let her go back to him. She says that she will never stop loving him. He promises her that he won t and they make love again. Joshua tells Elizabeth that he wants to take Alison as his bride. Elizabeth tells Joshua that she is not going to let him take her. She wants to stop him from taking Alison. Rafe tells Alison that he has a vision of a city of tortured souls, a master vampire race, and his vampire bride. He will never let that happen. He can t let Joshua get to her. He is going to keep her away from Joshua and all his goons. He wants her to be his bride and she will be his wife. She will never love him, Rafe says.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Joshua**
	• Who does Frank try to kidnap?	**Allison**
	• Who tries to kidnap Allison?	**Joshua**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**vampire bride**

Table 17: Examples of generated summaries and automatic QA evaluation for different content selection methods. In all model variants, we only consider the textual modality (i.e., BART with adapter tuning).

Model	Summary	
Gold	Joshua tells Elizabeth he wants to turn Allison and demands she help ease Allison into her new life as his wife. Elizabeth tells Joshua she will kill him before she allows him to hurt Allison. Livvie is able to fend off her need to feed while she and Caleb make love. Frank searches for Allison. When Frank attempts to kidnap Allison from Rafe, he discovers that it really is Lucy and I an in disguise. Allison and Rafe reappear in Caleb s cave.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Rafe**
	• Who does Frank try to kidnap?	**Allison**
	• Who tries to kidnap Allison?	**Frank**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**wife**
Text-only (CS)	Rafe tells Alison that he will never let Joshua take her for his bride, but she tells him that she has no choice in the matter. Elizabeth tells Joshua that she will not stand by and allow him to take her daughter. Joshua tells Elizabeth that he is going to eas e Alison into her new lifestyle as his wife. Elizabeth says that she is not going to let her daughter suffer the kind of nightmare that she lived. She will kill Joshua before he is even that close to turning her. Alison tells Rafe that she thinks this is a little extreme, that is all. Rafe says he will not let Joshua get to her. He promises to keep her away from Joshua and all his goons. Caleb tells Livvie that she doesn t need to feed. He tells her that he can t make love to her the way she wants to. She tells him she can t turn him back, but he tells her he can. He says that he loves her and that he wants to make her his bride.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Joshua**
	• Who does Frank try to kidnap?	**Joshua**
	• Who tries to kidnap Allison?	**Rafe**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**wife**
H-3D (CS)	Livvie tries to convince Caleb to let her take the risk of biting him, but she is afraid that she won t be able to do it. Joshua tells Elizabeth that he wants Alison to be his bride. Elizabeth is shocked when she finds out that Joshua wants to take Alison away from Rafe. Elizabeth tells Joshua that she will find a way to stop him from taking Alison. Rafe tells Alison that he has a vision of a city of tortured souls, a master vampire race, and his vampire bride. He tells her that he can make a perfect bride for her. Alison tells Rafe that she doesn t want to leave her family, but Rafe assures her that she is not going to leave them. Frank tells Ian that he is going to have to tell his boss that his mission didn t work.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Rafe**
	• Who does Frank try to kidnap?	**Allison**
	• Who tries to kidnap Allison?	**Rafe**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**vampire bride**
LED	Caleb and Livvie try to convince Caleb that they can be together again, but Caleb refuses to believe that he can t make love to Livvie. Rafe and Alison agree that they will be together, but Rafe tells Alison that he will never be able to make love with her again. Elizabeth tells Joshua that she wants him to turn Alison into a vampire bride. Joshua tells Elizabeth that he wants her to be his wife, but Elizabeth tells him that she is not going to do that. Joshua says that he is going to make Alison a vampire, and he will be the one true love of his life. Elizabeth says that she will never let Alison suffer the kind of nightmare that she lived, and she will make her life as a vampire. Joshua and Elizabeth argue about how much she wants to be a vampire and how much he wants to help her. Elizabeth asks Joshua if he s going to help Alison, but he says he will not.	
QA pairs	• Who does Frank try to kidnap Allison from?	**Caleb**
	• Who does Frank try to kidnap?	**Caleb**
	• Who tries to kidnap Allison?	**Rafe**
	• Who can fend off her need to feed while she and Caleb make love?	**Livvie**
	• Who tells Joshua she will kill him before she allows him to hurt Allison?	**Elizabeth**
	• Who tells Elizabeth he wants to turn Allison into his wife?	**Joshua**
	• What is Allison s new life?	**vampire**

Table 18: Examples of generated summaries and automatic QA evaluation for different models. Here we compare our Hierarchical3D model (H-3D) with state-of-the-art textual summarizers (i.e., LED).