End-to-End Multimodal Representation Learning for Video Dialog

Huda Alamri
Georgia Institute of Technology
halamri3@gatech.edu

Anthony Bilic
Georgia Institute of Technology
abilic3@gatech.edu

Michael Hu
Georgia Institute of Technology
mhu93@gatech.edu

Apoorva Beedu
Georgia Institute of Technology
abeedu3@gatech.edu

Irfan Essa
Georgia Institute of Technology
irfan@gatech.edu

Abstract

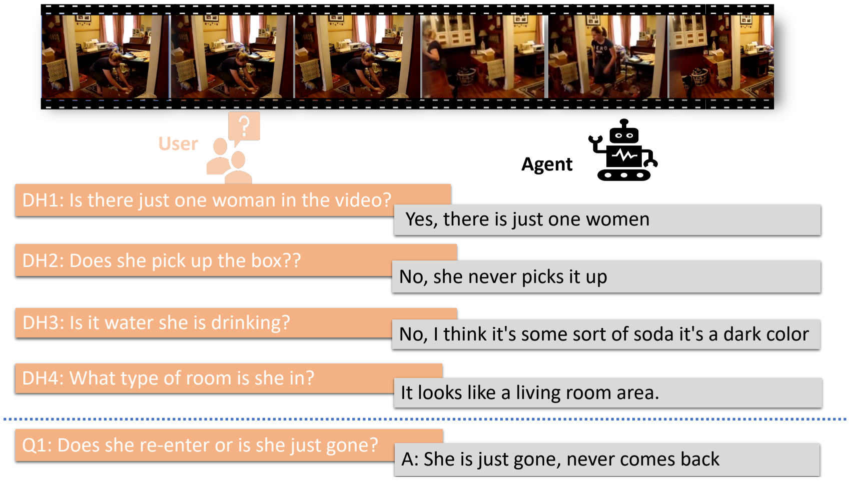
Video-based dialog task is a challenging multimodal learning task that has received increasing attention over the past few years with state-of-the-art obtaining new performance records. This progress is largely powered by the adaptation of the more powerful transformer-based language encoders. Despite this progress, existing approaches do not effectively utilize visual features to help solve tasks. Recent studies show that state-of-the-art models are biased towards textual information rather than visual cues. In order to better leverage the available visual information, this study proposes a new framework that combines 3D-CNN network and transformer-based networks into a single visual encoder to extract more robust semantic representations from videos. The visual encoder is jointly trained end-to-end with other input modalities such as text and audio. Experiments on the AVSD task show significant improvement over baselines in both generative and retrieval tasks.

1 Introduction

The goal of the video-based dialog task is to answer questions about a dynamic scene presented in the video. More precisely, given a short video clip and multiple rounds of questions and answers about the video, the model should provide an accurate response to a follow-up question. An example of this is shown in 1, where a model is presented with a short video and a conversation about it. When the model is asked a follow-up question: *"Did she re-enter the room?"*, to provide an accurate answer, the model has to acknowledge that the person "she" refers to the "woman" mentioned in the previous utterances. The model also has to identify the action "re-entering the room" from the actions in the video. This video-based dialog task represents a challenging multi-modal learning problem that serves as a test bed for video and language representation learning. Advances in this research field influences a wide range of applications, including providing road assistance for autonomous vehicles [14], helping visually impaired individuals to understand their surroundings, and navigating through a very long video etc.

Success in this multi-modal learning task hinges on tackling four main challenges: *(i)* extracting strong visual representations; *(ii)* extracting strong textual representations; *(iii)* effectively combining

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



DH1: Is there just one woman in the video?

Yes, there is just one women

DH2: Does she pick up the box??

No, she never picks it up

DH3: Is it water she is drinking?

No, I think it's some sort of soda it's a dark color

DH4: What type of room is she in?

It looks like a living room area.

Q1: Does she re-enter or is she just gone?

A: She is just gone, never comes back

Figure 1: In video dialog task, the model is presented with a short video, a dialog about the video, and a follow-up question. The goal is to correctly answer the question conditioned on the audio-visual cues and the dialog history (DH).

both features with other modalities (audio, when available); and finally, *(iv)* generating an accurate response in natural language. While the task has received considerable interest from the community, current work largely focuses on obtaining strong textual and visual representations independently and combining the features [3, 23, 34, 18, 17], while the knowledge and cues from the video-text association have not been extensively explored. This was investigated by Liu *et al.* [26], who demonstrated that most models are biased towards the textual information, while visual features not contributing substantially towards performance. This study argues that using the visual features extracted from frozen 3D-CNN networks learned from action recognition data, without the added knowledge about the corresponding text association, i.e. the questions, result in reduced performance compared to joint training with both modalities.

Our work addresses this limited utilization of visual information in the video-based dialog task by making the models more visually aware. First, a 3D-CNN network extracts local temporal features from the input video, which is then passed to a transformer based visual encoder network that generates contextual representation through self-attention mechanism. These visual features are then effectively combined with text and audio features to generate a best response for the input video and question. These multiple modules form one unified framework that is trained end-to-end which enables the model to generate stronger latent representations. Experiments on the video-based dialog task AVSD show that our model learns a stronger joint visual-textual features which contribute significantly to its performance. Through several baselines, we show how recent methods pre-extract visual features and improve the vision-based language tasks due to the strong performance of the language models (e.g.,BERT and GPT2). On the contrary, our framework is designed to use standard architectures to emphasize that joint learning of visual and textual information is vital for the video-dialog task.

The contributions of our work are as follows:

- We propose a new framework for video-based language understanding and generation tasks. This multi-modal framework effectively learns contextual representation using strong visual features from video, and through self-attention.
- Our framework is flexible and can use any number of modalities and different encoders for these inputs. We show ablations on using Audio in addition to Text and Video modalities in 4.
- We also show the effectiveness of joint training on the retrieval task with a simpler framework. We provide extensive experiments and detailed analysis of both generative and retrieval tasks in the AVSD dataset are provided.

2 Related Work

Video and language understanding has been extensively investigated due to the wide range of potential applications in human-computer interactions. Tasks such as video captioning [44, 49, 5], video question-answering [19, 48, 31, 20], and video dialog [1, 13, 24] study the complex interplay between the vision and natural language modalities. In the case of video question-answering, effective performance depends on extracting strong visual representation for the input video and efficiently fusing it with the associated text. For video dialog, Alamri *et al.* [1] introduced the Audio-Visual Scene Aware Task (AVSD) as a multi-modal learning problem, the objective of which is to answer a question based on a short video, with an associated audio and a dialog history. The task supports a discriminative setting, where the model ranks a list of candidate answers [1, 29], or a generative setting, where a decoder is trained to auto-regressively generate an answer [23, 11].

Self-attention models, known as transformers [42], have been very successful at generating deep contextual linguistic representations. They are generally pre-trained with self-supervised learning on very large unlabelled text corpora, and subsequently fine-tuned on downstream tasks. They deliver state-of-the-art results for several natural language understanding and generation tasks [42, 33, 32, 8]. In our work we utilize a pre-trained BERT[8] model to encode the input question and the dialog history.

Inspired by this success, a large body of work has adapted self-attention models to multi-modal learning, including image question answering [27, 7, 22, 21, 41], image dialog [7], video question answering [39, 38, 39], and video dialog [23, 3, 17]. In general, these approaches can be categorized into single-stream and two-stream networks.

In the two-stream approach, each modality is independently encoded by a transformer-based network, and information is fused through concatenation or cross-attention [41, 27]. In the one-stream approach, Li *et al.* [22], Su *et al.* [37], and Li *et al.* [29] utilize a unified transformer network where video and text tokens are combined as one sequence.

In the two-stream approach, the visual features and the text features extracted using modal specific encoders then fused jointly via transformer-based encoder Luo *et al.* [28]. This study builds on the proposed model in [28] and extends it in two ways: first, a 3D-CNN network is added to the backbone visual encoder. Second, an audio transformer-based encoder is added to learn a representation from the audio signal, which is combined with the other modalities via a cross-encoder, and the different encoders and the decoder are jointly trained in an end-to-end fashion. The experiments demonstrate the benefits of this approach.

Le. H. *et al.* proposed a multimodal transformer network with query-attention [17]. Zekang et al. *et al.* [33] utilized a pretrained GPT2 model and extended it to learn joint audiovisual and text featuring by training the model on multitask learning objectives [23]. Cherian A. *et al.* extend the audio-visual transformer by adding student-teacher learning [35]. While all these approaches for video dialog tasks have achieved promising improvements, the utilization of visual features remains limited. All the approaches rely on pre-extracted visual features from 3D-CNN networks with no further fine-tuning or training. This has resulted in models that do not fully capture the multimodal nature of the task [26]. In contrast, this model designed in this study also updates the visual extractor (a 3D-CNN) in an end-to-end fashion, which leads to the improved learning of visual features tailored to the video question answering task.

3 Method

This section introduces the framework for the video-based dialog task. It presents the different modal-specific encoders, pre-processing of the input modalities, training objectives, and the evaluation process.

3.1 Task Formulation

Given an input video $V{=}(V_1, \ldots, V_i, \ldots, V_n)$, where V_i is the i^{th} frame sampled from the video, a dialog history $DH_t{=}(C, (Q_1, \mathit{Ans}_1), \cdots, (Q_{t-1}, \mathit{Ans}_{t-1}))$, where C is the video caption and $(Q_{t-1}, \mathit{Ans}_{t-1})$ corresponds to a question-answer pair at round $t-1$, and audio A (see Figure 1), the task is formulated such that, given a follow-up question Q_t, the model must generate a response R_t considering input features: V, $DH_{1:(t-1)}$, A, and Q_t:

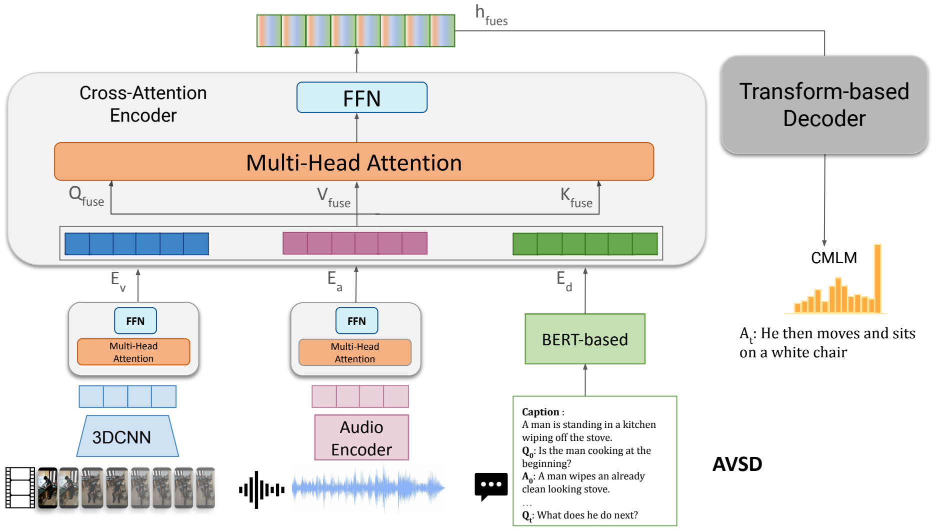

h_fues
Cross-Attention Encoder
FFN
Multi-Head Attention
Q_fuse
V_fuse
K_fuse
Transform-based Decoder
CMLM
A_t: He then moves and sits on a white chair
E_v
E_a
E_d
FFN
Multi-Head Attention
FFN
Multi-Head Attention
BERT-based
3DCNN
Audio Encoder
Caption :
A man is standing in a kitchen wiping off the stove.
Q_0: Is the man cooking at the beginning?
A_0: A man wipes an already clean looking stove.
…
Q_t: What does he do next?
AVSD

Figure 2: The proposed model consists of **Visual encoder** that receives sequences of frames and generates an embedding E_v, **Text Encoder** receives text tokens and generates an embedding E_d, **Audio Encoder** generates an embedding E_a, **Multi-Modal Encoder** fuses these embeddings and jointly train the encoders end-to-end.

$$P(R_t|V, A, DH_{1:(t-1)}, Q_t; \theta) = \prod_{j=0}^{t-1} P(R_j|V, A, DH_{1:j-1}, Q_t; \theta) \tag{1}$$

and train to minimize the cross entropy loss:

$$\mathcal{L}(\theta) = -\log P(R_t|V, A, DH_{1:(t-1)}, Q_t; \theta) \tag{2}$$

where θ comprises of the trainable network parameters.

3.2 Model Architecture

A general overview of the proposed model is presented in Figure 2. It consists of several multi-stream modal specific encoders to extract the initial features, followed by a self-attention network that applies self-attention encoders to generate the contextual representations followed by a transformer-based encoder that generates the final multimodal embedding via cross-attention mechanism. This is then passed to an auto-regressive decoder to generate an open-ended response.

3.2.1 Text Encoder

All the text inputs: DH, C, Q and Ans are concatenated to form a single long string. Following Devlin *et al.* [9] all the words are tokenized using the Word Piece tokenizer [45] to obtain a token sequence $t = \{t_i | i \in [1, n]\}$; where t_i is the i-th token, and n is the length of the language token sequence. *[CLS]* token is added at the beginning of the input sequence, and *[SEP]* is used to separate each sentence (the sentence is either a question or an answer). The processed tokens are then fed to a BERT-based uncased model [9] to generate a text embedding $E_d \in \mathbb{R}^{n \times d}$, where d is the hidden size of the final self-attention layer of BERT.

$$E_d = BERT(t) \tag{3}$$

3.2.2 Visual Encoder

Initially, a sequence of frames $V_n = \{v_j | j \in [1, m]\}$ was subsampled at 16 fps and cropped to 224 x 224. V_n is fed into a 3D-CNN network to extract the temporal features. We used *I3D* network [4] pretrained on ImageNet for the encoder. We extracted global average pooled features from different inception blocks such as *Mixed*$_4$ and *Mixed*$_5$ with dimensions m x d. Finally, a visual transformer-based encoder applies self-attention over these features f_v and generates visual embeddings E_v. The visual encoder consists of $N = 6$ layers of Multi-Head Attention (*MHA*) and Feed-Forward Network (*FFN*).

$$f_v = \mathit{I3D}(V_n), \tag{4}$$
$$E_v = \mathit{FFN}(f_v) + \mathit{MHA}(f_v). \tag{5}$$

3.2.3 Audio Encoder

To process the audio input, m-dimensional features were first extracted using a *VGGish* [10] network. Similar to the visual features, these were then fed into a transformer-based encoder to extract a contextual representation. Unlike the visual features, the audio CNN network was not fine-tuned.

$$f_a = \mathit{VGGish}(A_m), \tag{6}$$
$$E_a = \mathit{FFN}(f_a) + \mathit{MHA}(f_a). \tag{7}$$

3.2.4 Cross-Attention Encoder

Finally, to generate the multimodal representations, we adapted a cross-attention encoder proposed in [28] and extended it to learn the interdependencies between the three different modalities. Given the visual E_v, audio E_a and dialogue embeddings E_d, the encoder fuses them into one sequence and applies cross-attention mechanism over them [28]. The cross-encoder consists of $N = 6$ *MHA* layers followed by a Feed Forward Network. Finally, a transformer-based auto-regressive decoder is trained to generate responses given the multi-modal representation h_{embd}.

$$H_{\text{fuse}} = ([E_v; E_a; E_d]), \tag{8}$$
$$h_{\text{embd}} = \mathit{FFN}(H_{\text{fuse}}) + \mathit{MHA}(H_{\text{fuse}}). \tag{9}$$

3.2.5 Training and inference

The model is trained by optimizing for two objectives losses introduced in [28], namely the Conditioned Masked Language Modeling *CMLM* and the Decoder Reconstructive Loss.

For ***CMLM***, $\%15$ of the input text tokens were masked with *MASK* special token and the model was trained to predict the masked tokens conditioned on h_{embd}.

For the **Decoder Reconstructive Loss**, at each iteration, the decoder receives the encoded embeddings and generates one answer token $\hat{y}_{i+1}$ that is conditioned on the multi-modal fused output and previous generated word $\hat{y}_i$. At inference time, $\hat{y}_{i+1}$ with the highest score was chose.

$$\hat{y}_{i+1} = argmaxP(y_{i+1} = y|\hat{y}_i, h_{embd}), \tag{10}$$

4 Experiments

4.1 Dataset and evaluation metrics

We evaluated our framework on the Audio-Visual Scene-Aware Dialog (AVSD) dataset [1]. It comprises of dialogs grounded in human-based action videos and videos from the Charades [36] dataset. Each dialog consists of a video caption and 10 rounds of questions and answers about the events in the video. In total, there are 7,659, 1,787, and 1,710 dialogs in the train, val and test sets respectively.

The results on the DSTC-test set are reported using the common natural language generation evaluation metrics including:**BLEU** [30], **METEOR** [2], **ROGUE-L** [25], and **CIDEr** [43]. The test set has only one correct answer for each question.

4.2 Preprocessing:

Dialog History: For the dialog input, we use up to 3 turns of dialog history with a maximum length of 100 words, which was generally sufficient for 3 rounds of dialog history plus the question.
Video-Audio features: We extracted 1024-d feature from ***Mixed***$_{5c}$ and ***Mixed***$_{4c}$ layers. For comparison with the AVSD baselines [1], we pre-extracted the visual features using I3D [4] trained on the ImageNet dataset and used the 1024-d output from the ***Mixed***$_{5c}$ layer for the baseline. For the Audio modality, we use pretrained 1024-d features from *VGGish* [10]. This encoder is not fine-tuned on the AVSD dataset.

4.3 Training

We use the Adam optimizer [15] with a learning rate of $5e^{-5}$ and batch size of 64. Training was done using 8 RTX-6000 GPUs. Early stopping and checkpoint that achieved the best performance on the validation set was selected.

5 Results and Analysis

In this section, we first perform a detailed analysis on the generative task. For a more well-rounded understanding of the contribution of visual features, several ablations on the retrieval task were performed, – where the answer is retrieved from a pool of candidate options. Finally, we demonstrate the performance of the proposed method via qualitative evaluation.

Table 1: Model performance on the AVSD test for the generative task. * includes Audio, † includes summary.

Method	BLEU2↑	BLEU3↑	BLEU4↑	METEOR↑	ROGUE-L↑	CIDEr↑
DGR* (2021)	–	–	0.357	0.267	0.553	1.004
JST*†(20219)	–	–	0.406	0.262	0.554	1.079
VideoGPT2*† (2020)	0.570	0.476	0.402	0.254	0.544	1.052
MTN † (2019)	0.242	0.174	0.135	0.165	0.365	**1.366**
JMAN (2020)[6]	0.521	0.413	0.334	0.239	0.533	0.941
Le H. et. al.(2021)[16]	0.577	0.476	0.398	0.262	0.549	1.040
TimeSformer *† (2022) [47]	0.572	0.477	0.403	0.255	0.547	1.049
Ours + Audio modality*	0.587	0.483	0.401	**0.271**	0.565	1.155
Ours	**0.592**	**0.493**	**0.415**	0.269	**0.569**	1.159

5.1 Results on the Generative task

We compare our results with [17, 23, 35]. VideoGPT [23] uses GPT2[33], a pretrained generative encoder that is known to outperform BERT[8] model that we adapted. JST and MTN are also self-attention based models, however they do not finetune the visual backend network to AVSD dataset and feed pre-extracted visual features. Table 1 Shows that our model outperforms these models across the different evaluation metrics, achieving a gain in BLUE2 (0.592 –>0.570), BLUE3 (0.493–>0.476), BLUE4 (0.415–>0.406), METEOR (0.269 –>0.267) and ROUGEL (0.569 –>0.356). For CIDEr, although our method underperforms to the MTN method, the latter uses a much larger model and more textual input (Summary, in addition to caption, and dialog history). These results indicate that the joint training improves the model‘s utilization of the visual features, and with only a slight increase in memory and time cost, performs better or comparable results to models with deeper networks. By using standard architectures, we highlight the gains due to the textual-visual association rather than stronger language encoder that is not visually-aware. We would like to reiterate that the novelty of the proposed work lies in the approach taken in learning the joint features, and the performance improvement achieved speaks to that.

Role of Audio Modality:

In Table 1, we show methods that use Audio with *. When compared to our method that uses Audio and Text inputs, using Audio does not show a significant improvement. This is because, for AVSD dataset specifically, the audio has sounds without any dialog. However, for completeness, and generalisation to other datasets, we have included the results in the table.

5.2 Results on the Retrieval task

To establish that visual information aids effective performance in the AVSD task, we further evaluate our proposed approach on the retrieval setting. In the retrieval setting, the model is given the same inputs as the generative task but tasked with retrieving the correct answer from a pool of candidate answers by outputting a ranking. This settings allows for direct evaluation of the encoded modalities without the decoder performance.

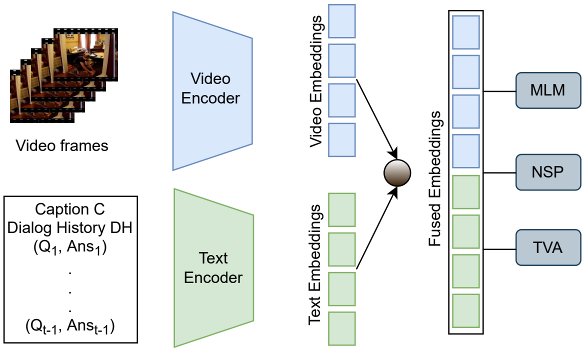


Figure 3: Retrieval task: The classification model consists of Visual encoder that receives sequences of frames and generates an embedding E_v,Text Encoder receives text tokens and generates an embedding E_d,Dialog Encoder fuses these embeddings and jointly train the encoders end-to-end.

For this purpose, we design a much simpler framework as shown in 3, where the video embeddings from I3D and text embeddings from the BERT model are concatenated and optimized for the following objectives: Masked Language Model loss (L_{mlm}), Next Sentence Prediction loss (L_{nsp}), and text-video alignment loss (L_{vta}). The retrieval model is also trained jointly, thus learning from the visual-text association. The training objectives are detailed in the Appendix 6.

Table 2: Model performance on the AVSD dataset. XXX$_{\text{ft}}$ refers to finetuned models, XXX$_{\text{no-ft}}$ to non-finetuned models. ↑ implies higher the score the better, ↓ implies, lower the score the better.

Input	**Text Encoder**	**Vid Encoder**	↑ **MRR**	↑ **R@1**	↑ **R@5**	↑ **R@10**	↓ **MR**
DH	LSTM	-	50.40	32.76	73.27	88.60	4.72
	BERT	-	69.71	56.93	86.18	92.93	5.07
DH + V	LSTM	I3D$_{\text{no-ft}}$	53.41	36.22	75.86	89.79	4.41
	LSTM	S3D$_{\text{no-ft}}$	53.57	36.49	75.64	89.82	4.45
	LSTM	I3D$_{\text{ft}}$	54.28	37.12	76.62	90.23	4.33
DH + V (Ours)	BERT$_{\text{ft}}$	S3D$_{\text{no-ft}}$	71.32	59.51	86.92	**95.22**	4.89
	BERT$_{\text{ft}}$	S3D$_{\text{ft}}$	**77.28**	**67.28**	**90.39**	94.87	**4.18**

Evaluation Metrics We report the retrieval metrics: R@1, R@5, R@5, as well as the Mean Rank (MR), and Mean Reciprocal Rank (MRR). Ideally the ground truth answer is ranked first.

Performance of the language encoders: Table 5.2 summarizes the results of the evaluation phase. For the text encoders, we note that BERT significantly outperforms the LSTM encoder, achieving 69.71 MRR, which is a 19% absolute improvement over the LSTM-based encoder. This development was anticipated as Transformer-based encoders, such as BERT, benefit from pre-training with a large text corpora on several proxy tasks. This can generate a rich contextualized representation that assists the model to understand linguistic input. However, we would like to note that even when a simple language encoder such as LSTM is used, jointly training the visual encoder results in a significant performance improvement.

Effect of different visual encoders: We measure the effect of utilizing different 3D-CNN networks to extract the visual representations, and observe that they perform comparably, achieving 53.41 MRR for the I3D features, and 53.57 for S3D (5.2). This indicates that alterations to the visual encoder do not improve the model performance, as the actual joint training for the network was the chief factor for the performance.

Joint training is effective: Combining the visual features with the fine-tuned language BERT encoder - BERT$_{\text{ft}}$ $+$ S3D$_{\text{no-ft}}$, outperforms the language only model with an increase of 1.6% on MRR. This modest improvement when adding visual features has been the main trend in video-based dialog systems [1, 11, 39, 12]. Finally, the model that is jointly fine-tuned on both modalities achieved the best performance across all metrics, with a 6% increase in MRR.

5.3 Effect of the number of fine-tuned blocks.

In Table 5.3 we detail the impact of fine-tuning several inception blocks on our visual encoders. The S3D network comprises three convolution layers, followed by five inception blocks. The depth and network architecture resulted in additional trainable parameters. The aim was to learn the effect of conducting fine-tuning across more layers, i.e. inception blocks, to ascertain whether this would allow the model to generate better visual features for the task. Table 5.3 shows the model's performance when fine-tuning different inception models.

Table 3: Performance by finetuning different inception blocks from the visual encoder

	Retrieval Mode					Generative Mode					
Trained Inception Blocks	↑MRR	↑R@1	↑R@5	↑R@10	↓MR	↑BLEU2	↑BLEU3	↑BLEU4	↑METEOR	↑ROGUE-L	↑CIDEr
S3D$_{\text{no-ft}}$	53.41	36.22	75.86	89.79	4.41	0.58	0.488	0.407	0.268	0.561	1.115
S3D$_{\text{Mixed5}}$	**77.21**	**67.20**	90.22	**95.06**	**4.15**	**0.592**	0.492	0.413	0.267	0.563	1.134
S3D$_{\text{Mixed4,Mixed5}}$	76.88	66.72	**90.39**	94.77	4.48	**0.592**	**0.493**	**0.415**	**0.269**	**0.569**	**1.159**

5.4 More frames are more informative:

We also conducted an experiment in which we varied the size of the sampled frames. The question we are seeking to answer was: *how many video frames are sufficient to answer the input question?* We trained the model using sampling frame sizes: 6, 16, 32 and 40. As presented in Table 4, we can see that the model performance improved significantly when trained on larger sampling rates, concluding that the model benefits greatly from additional visual features when trained jointly on downstream tasks. We see a small drop in performance at frame rate of 40 frames. We believe this is because the pre-trained model was trained with 30 frames per video sequence, and an increasing the number of frames results in redundant data.

5.5 Dialog history is helpful:

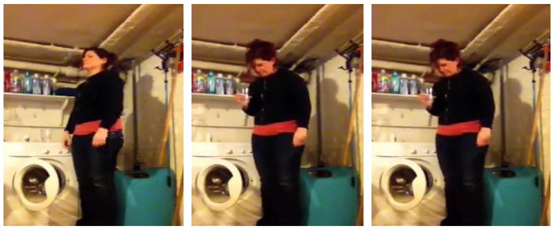
We evaluated the effect of the length of the dialog information. We tested the model performance in the first round -Round$_1$, where there were no prior dialog utters, and in Round$_3$, where there were 2 previous dialog utters, increasing in Round$_5$ and Round$_{10}$. The results for each round are displayed individually in the AVSD test set in 5. As we can observe, the model performance improved from the first round, with 59.57 MRR, to the third round, which obtains 81.66 MRR. This was because the third round included information from previous rounds. As the dialog tends to become more generic and uninformative after the initial Q and As as seen Figure 4, we see a performance drop after the third round.

Table 4: Evaluation results for on the test set of AVSD Performance by total number of sampled video frames

Number of Sampled Frames	↑MRR	↑R@1	↑R@5	↑R@10	↓MR
6	46.38	31.15	64.70	78.87	8.46
16	74.90	64.11	89.19	94.47	4.63
32	**77.28**	**67.92**	**90.22**	**95.06**	**4.15**
40	77.21	66.20	89.62	94.82	4.46

Table 5: Evaluation of length of the dialog history on the Performance.

Dialog Round	MRR	R@1	R@5	R@10	Mean rank
1	59.57	46.39	75.70	87.71	4.49
3	**81.66**	**71.65**	**94.82**	**98.81**	**1.82**
5	77.21	67.20	89.62	94.82	4.46
10	62.52	52.96	74.46	78.77	18.34



Id	Question	GT_Answer
0	what is the person looking at in the beginning?	she is looking at the glass of water in her hands
1	Is she in the laundry room?	Yes, it looks to be in the basement of her home
2	Is she doing laundry at all	No, she sneezes and takes some medicine
3	Where does she get the medicine from?	A bottle on the washing machine
4	Does she put the bottle back on washing machine?	Yes, she does, then she drinks the water
5	Does she set the glass down?	She sets the glass down after sneezing to get the medicine
6	Does she ever move around?	No she stays in the same place
7	Does the video end with her drinking water	Yes she is drinking the water at the end.
8	Does she say anything at all	No she does not speak.
9	Are there any noises in video	Only her sneeze can be heard chips

Figure 4: Questions and answers in a typical dialog setting. We see that the first few questions are closely related to the video, but the later ones are very generic which makes it hard for the BERT model to train on.

6 Conclusion

In this paper, we proposed a new framework for a video-based dialog task. In our framework we optimized the learning from the visual input by jointly training the visual encoder end-to-end with different modalities like text and audio. Different generative and retrieval tasks showed that our training scheme generates a more rich multimodal representation and helps reduce the bias towards the textual information. We emphasize that **joint learning of visual and textual information is vital** for the video dailog task. Though there is an additional time cost and memory cost, our results show a significant improvement across all tasks and metrics, thus re-enforcing our belief that the finetuning the video encoders is crucial for the tasks. Future work will aim to extend our goal to pretrain our model using self-supervision tasks in raw unlabeled data then use the generated representations for more complex tasks such as video captioning, and video retrieval.

References

[1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[3] Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*, 2019.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] Ming Chen, Yingming Li, Zhongfei Zhang, and Siyu Huang. Tvt: Two-view transformer network for video captioning. In *Asian Conference on Machine Learning*, pages 847–862. PMLR, 2018.

[6] Yun-Wei Chu, Kuan-Yen Lin, Chao-Chun Hsu, and Lun-Wei Ku. Multi-step joint-modality attention network for scene-aware dialogue system. *arXiv preprint arXiv:2001.06206*, 2020.

[7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[11] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356. IEEE, 2019.

[12] Chiori Hori, Anoop Cherian, and Tim K Marks. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc8. In *DSTC7 at AAAI, 2019 Workshop*, 2019.

[13] Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. Video dialog via multi-grained convolutional self-attention context networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 465–474, 2019.

[14] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Hung Le, Nancy F Chen, and Steven CH Hoi. c^3: Compositional counterfactual constrastive learning for video-grounded dialogues. *arXiv preprint arXiv:2106.08914*, 2021.

[17] Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166*, 2019.

[18] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. Dstc8-avsd: Multimodal semantic transformer network with retrieval style word generator. *arXiv preprint arXiv:2004.08299*, 2020.

[19] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[20] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019.

[21] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[22] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[23] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, Cheng Niu, and Jie Zhou. Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog. *arXiv preprint arXiv:2002.00163*, 2020.

[24] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2476–2483, 2021.

[25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[26] Aishan Liu, Huiyuan Xie, Xianglong Liu, Zixin Yin, and Shunchang Liu. Revisiting audio visual scene-aware dialog. *Neurocomputing*, 2022.

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[28] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[29] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pages 336–352. Springer, 2020.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[31] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15526–15535, 2021.

[32] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[34] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12548–12558, 2019.

[35] Ankit P Shah, Shijie Geng, Peng Gao, Anoop Cherian, Takaaki Hori, Tim K Marks, Jonathan Le Roux, and Chiori Hori. Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning. *arXiv preprint arXiv:2110.06894*, 2021.

[36] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.

[37] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[38] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 3(5), 2019.

[39] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[41] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[43] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[44] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018.

[45] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[46] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 1(2):5, 2017.

[47] Yoshihiro Yamazaki, Shota Orihashi, Ryo Masumura, Mihiro Uchida, and Akihiko Takashima. Audio visual scene-aware dialog generation with transformer-based video representations. *arXiv preprint arXiv:2202.09979*, 2022.

[48] Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. Videoqa: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 632–641, 2003.

[49] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018.

Appendix

A Retrieval task: Implementation details

A.1 Text Encoder

To generate the text representation we utilize the BERT-based uncased model [9]. We feed the processed tokens to the text encoder to generate a text embedding $T_d \in \mathbb{R}^{n \times d}$, where d is the hidden size of the final self-attention layer of BERT.

We concatenate DH, C,C and Q to form a single long string. Next, we follow Devlin J *et al.* [9] and tokenize all the words using the Word Piece tokenizer [45] to obtain a token sequence $t = \{t_i | i \in [1, n]\}$; where t_i is the i-th token, and n is the length of the language token sequence. $< CLS >$ token is added at the beginning of the input sequence, and $< SEP >$ is used to separate each sentence (the sentence is either a question, or an answer). In addition to the embedding of these words, we add positional embedding, and segment embedding. For the segment embedding we followed [29] and added additional segment embeddings for the questions and answers, see Figure 5.

The hidden size of the model is 768 and the batch size is 16. For the dialog input, we used up to 3 turns of dialog history with a maximum length of 200 words.

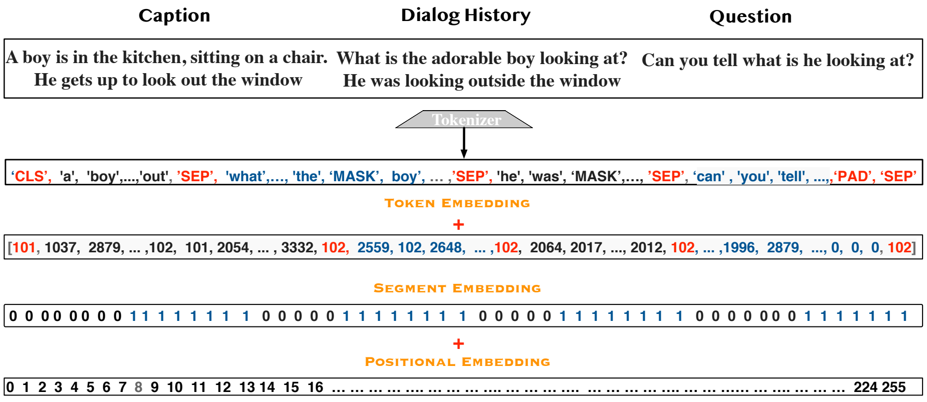


Figure 5: The final input for the text encoder is the sum of the token embedding, segment embedding and the position encoding

A.2 Visual Encoder

For the visual encoder, we use S3D [46], which is a 3DCNN built on the Inception network [40] as a backbone with separable temporal convolutions.

First, we subsample a sequence of frames $v = \{v_j | j \in [1, m]\}$ at the rate of m fps and 224 x 244 frame size. Next, we feed the frames sequence to the pretrained S3D and extract global average pooled features from different inception blocks such as $Mixed\_4$ and $Mixed\_5$. The extracted features, V_e have the dimension m x 1024. We present ablations over different frame rates in Section 5.4.

A.3 Training Objectives

We train our model jointly end-to-end by optimizing for the following objectives: Masked Language Model loss (L_{mlm}), Next Sentence Prediction loss (L_{nsp}), and text-video alignment loss (L_{vta}). MLM and NSP tasks are widely used to perform transfer learning and fine tuning of pre-trained Transformer-based models in new downstream tasks [9, 42]. In our work we adopted those tasks as follows:

- For the Masked Language Modelling task, we masked 10% of the final input tokens and replaced those tokens with a <MASK> token; and trained the model to predict the masked tokens from the surrounding ones.

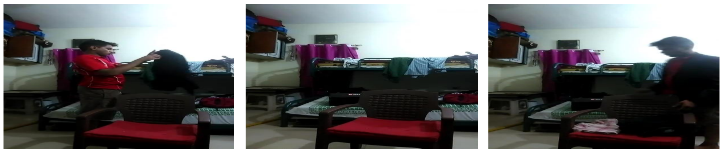
- In the Next Sentence Prediction task, given two sentences `A` and `B` a label **O**, we trained the model to give an output of **1** if `A` and `B` are related and should appear together, and output `0` otherwise. `A` is the concatenated input tokens: $C + DH_t + Q_t$, and `B` represents the ground-truth answer for a positive example. For a negative example, we randomly select the sample from the list of candidate answers.
- Text-video alignment task: After the visual embedding V_e is extracted using the aforementioned visual encoder, we apply a fully-connected layer to transfer V_e to the same dimensional space at T_e. Then we concatenate these representations to get the fused token embedding fused$_e$. Then we apply inner product between fused$_e$ and a candidate answer embedding a_e, and train the model with negative log-likelihood and *k*-negative samples with the weighted total losses: L_{mlm}, L_{nsp} and L_{vta}.

A.4 Qualitative results:

We show qualitative results from our retrieval model in Figure. 6 and Figure. 7. We see that our model is able to retrieve better answers compared to the video encoder is not finetuned to the task. As presented in these example, our best model ranks the correct answer as the top predicted answer more frequently than a the baseline model, with the same language encoder. Indicating that the model benefits largely from the joint end-to-end training of the visual and language encoders.

Question	**GT_Answer**	**Top_Predicted_Answer** BERT$_{\text{ft}}$ + S3D$_{\text{no-ft}}$	**Top_Predicted_Answer** BERT$_{\text{ft}}$ + S3D$_{\text{ft}}$
Is there only one man in the video?	Yes, he is the only person in the video	Yes, just the one man.	There is one person in the video.
In the beginning, is he looking at a book?	Yes, he is flipping though it before tossing it into the box	Yes it is, might be checking for something.	Yes, he is flipping though it before tossing it into the box
Do you think it is an instruction manual?	It may have been, it didn't look very thick	Yes probably. like a tall hair	It may have been, it didn't look very thick
Does he toss it in the blender box?	Yes, in the blender box	Yes, both the kitchen and the living room can be seen in the video.	Yes, in the blender box
What happens next in the video?	He starts wiping the island top off with a cloth	Yes, probably. like a tall hair	Yes, probably. like a tall hair
Is he using a cleaner or just a cloth?	He is only using the cloth	The is only using the cloth	The is only using the cloth
What happens after that?	Nothing. The video ends with him still wiping the island top off	Nothing. The video ends with him still wiping the island top off	Nothing. The video ends with him still wiping the island top off
So does he ever leave the room?	No, he stays in the kitchen	Yes, at the very end	Yes, at the very end
No one else enters the room?	No, there is no one else in the room	There is only one person	There is only one person
Does he ever look directly at the camera?	No, only at the book and the island top	No, only at the book and the island top	No, only at the book and the island top

Figure 6: Example 1: We show the answers generated by our model (Jointly trained BERT + Video Encoder) and compare it with the separately trained BERT + Video Encoder. The answers in green are the correct answers while red are incorrect answers generated by the models



	Question	GT_Answer	Top_Predicted_Answer	Top_Predicted_Answer
			$\text{BERT}_{\text{ft}} + \text{S3D}_{\text{no-ft}}$	$\text{BERT}_{\text{ft}} + \text{S3D}_{\text{ft}}$
0	Where does the video take place?	Looks like a bedroom maybe	It seems to start in a pantry.	Looks like a bedroom maybe
1	How many people are in the video?	There is just one person	One older gentleman is the person in the video	One older gentleman is the person in the video
2	What gender does the person in the video appear to be?	The person in the video is male	The person in the video is male	The person in the video is male
3	How old does he appear to be?	Not sure, maybe in his twenties	He looks young, id guess teens maybe	Not sure, maybe in his twenties
4	What happens after he enters the room?	He grabs a towel and some clothes and then goes off screen	He grabs a towel and some clothes and then goes off screen	He grabs a towel and some clothes and then goes off screen
5	Does he place the towel and clothes anywhere?	Not that i can see because he goes off screen with them	Not that i can see because he goes off screen with them	Not that i can see because he goes off screen with them
6	Does he ever come back on screen?	Yes he comes back without the stuff and then starts taking his jacket off	No he stays in that room for the rest of the clip	Yes he comes back without the stuff and then starts taking his jacket off
7	Where does he place his jacket?	He carefully places it on the bottom bunk of the bed	He is putting them on counters it looks like	He carefully places it on the bottom bunk of the bed
8	How does the video end?	He goes to his belt like he is going to take it off but he is kind of off screen and then it ends	He goes to his belt like he is going to take it off but he is kind of off screen and then it ends	He goes to his belt like he is going to take it off but he is kind of off screen and then it ends
9	Did you hear any sounds in the video?	No I don't hear anything in the video	No I heard no speaking	No I don't hear anything in the video

Figure 7: We show the answers generated by our model (Jointly trained BERT + Video Encoder) and compare it with the separately trained BERT + Video Encoder. The answers in green are the correct answers while red are incorrect answers generated by the models.