VX2TEXT: End-to-End Learning of Video-Based Text Generation From Multimodal Inputs

Xudong Lin[1] Gedas Bertasius[2] Jue Wang[2] Shih-Fu Chang[1] Devi Parikh[2,3]
Lorenzo Torresani[2,4]
[1]Columbia University [2]Facebook AI [3]Georgia Tech [4]Dartmouth

Abstract

We present VX2TEXT*, a framework for text generation from multimodal inputs consisting of video plus text, speech, or audio. In order to leverage transformer networks, which have been shown to be effective at modeling language, each modality is first converted into a set of language embeddings by a learnable tokenizer. This allows our approach to perform multimodal fusion in the language space, thus eliminating the need for ad-hoc cross-modal fusion modules. To address the non-differentiability of tokenization on continuous inputs (e.g., video or audio), we utilize a relaxation scheme that enables end-to-end training. Furthermore, unlike prior encoder-only models, our network includes an autoregressive decoder to generate open-ended text from the multimodal embeddings fused by the language encoder. This renders our approach fully generative and makes it directly applicable to different "video+x to text" problems without the need to design specialized network heads for each task. The proposed framework is not only conceptually simple but also remarkably effective: experiments demonstrate that our approach based on a single architecture outperforms the state-of-the-art on three video-based text-generation tasks—captioning, question answering and audio-visual scene-aware dialog.*

1. Introduction

Among the fundamental goals of AI is the development of conversational multimodal systems that can reliably perceive the real-world and communicate with humans in natural language. Progress in this area has been dramatically advanced in recent years by the introduction of large-scale benchmarks assessing the ability to interpret audiovisual information and translate this understanding to natural language. Prime examples include datasets for image or video captioning [10, 38, 51, 24, 56, 28], question answering (QA) [5, 13, 54, 58, 19, 36, 46, 26], as well as audio-visual dialog [11, 1]. In order to perform well on such benchmarks, the model must accomplish several goals: (1) extract salient information from each individual modality, (2) effectively combine the different cues to address the given query, and (3) generate and present the results in human-comprehensible text.

In this paper, we present VX2TEXT, a simple video-based approach that embeds these three steps in a unified, end-to-end trainable framework. Objectives (1) and (2) are accomplished by utilizing modality-specific classifiers to convert the semantics from each input signal into a common semantic language space, which enables the application of powerful language models to directly interpret multimodal content. Specifically, our approach takes the textual labels of the top classes predicted by each classifier pretrained on existing datasets [9, 14] and transforms them into word embeddings, using a pretrained language model [12, 40]. The benefit of this solution is that it opens up the possibility to carry out multimodal fusion by means of powerful language encoders such as T5 [40] without the need to design specialized cross-modal network modules [33, 29, 57, 35] or to resort to pretext tasks to learn to combine the different input signals [44, 55, 29]. Not only is such a design much simpler but it also leads to better performance compared to prior approaches.

In order to fulfill objective (3), we employ a generative text decoder [40], which transforms the multimodal features computed by the encoder into text, thus realizing the goal of generating results in human-comprehensible language. While prior multimodal works based on encoder-only architectures [44, 45, 33] are limited to operate in settings involving selection from a fixed set of text candidates, our generative approach can be used for open-ended sentence generation as, e.g., required in dialog applications. In addition, the use of a text decoder allows us to tackle different "video+x to text" problems (e.g., answering and generating questions, dialog, as well as captioning) with the same architecture, without having to design specialized network heads for each task.

We integrate these conceptually-distinct steps into a single architecture, which we train end-to-end. To achieve this, we adopt a differential tokenization on continuous modalities (e.g., audio or video) which renders the entire model—including the modality-specific classifiers—trainable with respect to the final objective. Our experiments demonstrate that our unified framework trained end-to-end produces significant performance gains over separately learned modules. Our VX2TEXT based on a single architecture trained in a generative fashion without any multimodal pretext pretraining outperforms the state-of-the-art on three different text-generation tasks—captioning, QA and dialog.

2. Related Work

Significant progress has been made in the area of vision and language, especially in the design of multimodal conversational agents that interact with humans in natural language, e.g., for question answering (QA) [5, 13, 54, 58, 19, 36, 46, 26] and audio-visual dialog [11, 1]. Several approaches have been introduced for these tasks [34, 3, 23, 42, 16, 41, 25, 52, 20, 27, 30, 55, 17, 31].

For example, Shah *et al.* [42] have proposed to leverage the cycle-consistency of question answering and question generation to improve the robustness of image question answering models on rephrased questions. Differently from our approach, the actual question and answer sentences are not decoded. Yang *et al.* [52] have explored an encoder-only model using multimodal fusion of BERT representations and visual features for video QA [26]. Being a discriminative approach, it is limited to selecting from the provided answer choices. Le *et al.* [25] proposed a multimodal attentional generative model, which fuses information from texts and audiovisual features and generates responses for audiovisual scene-aware dialog. While this and a few other recent works [55] have leveraged decoders for text-generation from multimodal inputs, we believe we are the first to empirically demonstrate via systematic ablations the performance improvements achieved by means of generative learning with decoding, compared to discriminative learning applied to the same encoder model. Furthermore, we note that the networks proposed in [25, 55] include specialized cross-modal blocks which, as noted above, approach the task quite differently from our method. Experimental comparisons to these prior works show the superior performance of our design.

There is also a family of multimodal transformer models [44, 55, 29, 45, 33, 31] leveraging pretext tasks inspired by the language domain [12, 39, 40]. These works rely on costly pretext training on large-scale datasets to learn multimodal representations. Conversely, our VX2TEXT can perform multimodal fusion in a unified language space and does not require multimodal pretext training.

We note that we are not the first to propose using labels of categories recognized from the audiovisual channels as input to language models. For example, detected object labels have been employed for image captioning [53, 4] and also video QA [26]. However, differently from these prior works, we adopt a differentiable tokenization on continuous modalities which makes the entire model trainable end-to-end with respect to the final objective. Our experiments demonstrate the performance benefits of our approach.

3. Technical Approach

Our goal is to design a unified framework that can generate open-ended text from video and accompanying modalities, e.g., audio, speech, or dialog history. We are specifically interested in tasks such as video captioning, question answering and audio-visual scene-aware dialog.

Formally, let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$ be a multimodal input sample, where $\mathbf{x}_m$ denotes the m-th modality. We specify the task that we want our model to address using a special task token $t \in \{Answer, Caption, Dialog, ...\}$. Our goal is then to train a model $\mathcal{F}(t, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M; \mathbf{W})$ that generates a sequence of text tokens $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$ representing the output for task t. $\mathbf{W}$ denotes the trainable parameters. Depending on the task, our generated text may be in the form of answers, questions, interactive responses in a dialog, or captions.

At a high-level, our approach can be summarized in three steps. First, we leverage pretrained modality-specific classifiers to obtain most probable category predictions for each modality. We then embed the textual names of the predicted categories into a semantic language space via our proposed differentiable tokenization scheme, which enables end-to-end training of the whole system including the modality-specific classifiers. Finally, we employ a generative encoder-decoder language model [40] for mapping the embedding vectors from the multiple modalities into free-form text. This allows us to reformulate different "video+x to text" problems as a single sequence-to-sequence task. We now present each of these steps in more detail.

3.1. Differentiable Tokenization

Most prior methods [25, 52, 29] rely on extra cross-modal fusion modules for combining input signals from different modalities. This renders the integration of different modalities burdensome and computational costly. Instead, we propose to perform multimodal fusion by mapping the different input signals into a common semantic language space through a simple scheme. We first leverage modality-specific classifiers trained to predict a large set of categories over predefined language vocabularies. These include video models trained to recognize a large collection of actions [9], or audio classifiers distinguishing a broad set of sound categories [14]. Afterwards, we can utilize existing language embedding models to map the top textual categories predicted by each modality-specific classifier into a common semantic language space.

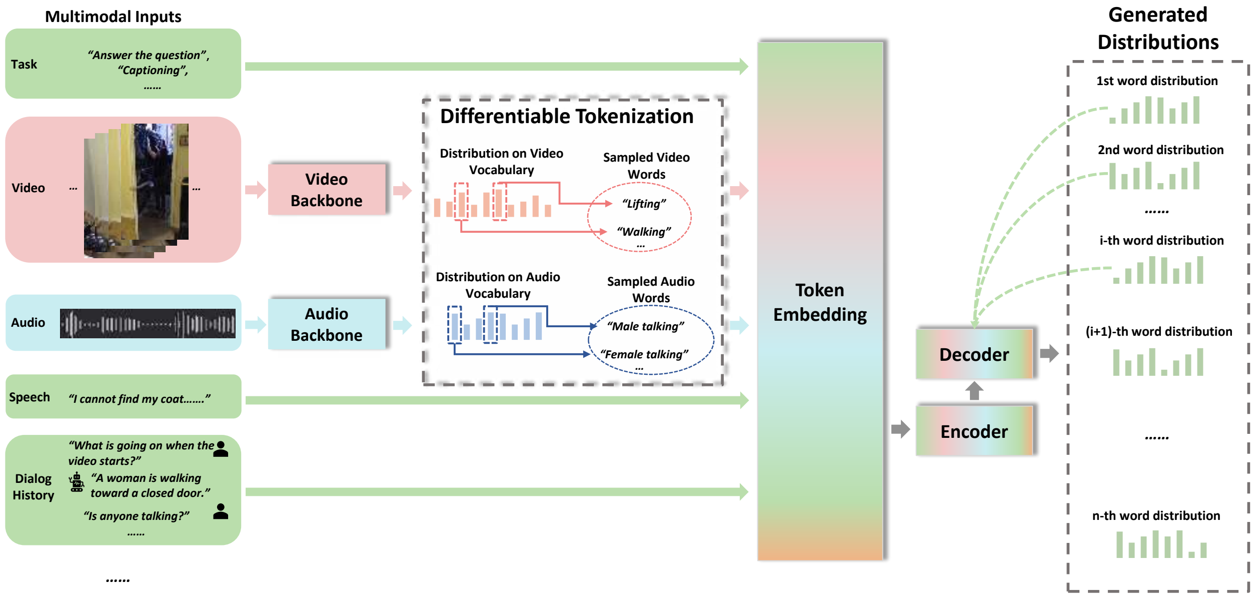


Figure 1. Illustration of our proposed framework. VX2TEXT receives as input a task specifier, and video with accompanying modalities, such as audio and speech. Each modality is converted into a set of tokens by means of modality-specific classifiers and a differentiable tokenization scheme that enables end-to-end training. Finally, an encoder-decoder architecture performs multimodal fusion in the language space and generates as output open-ended text addressing the given task.

Although conceptually simple, this approach has a few weaknesses. First, the pretrained modality-specific classifiers may not generalize to the target data. Second, the selection of the top categories from each classifier is not differentiable and thus prevents us from finetuning the modality-specific classifiers with respect to our target task. To address these limitations, we propose a differentiable tokenization scheme, which enables end-to-end training of the whole system including the modality-specific classifiers.

Let us denote with $\{\mathcal{N}_1, \mathcal{N}_2, ..., \mathcal{N}_M\}$ a set of modality-specific networks. For each modality m we use a network model $\mathcal{N}_m$ pretrained for a classification task on a predefined category space $\mathcal{C}_m = \{1, ..., C_m\}$. Let $p_m(c|\mathbf{x}) \in [0, 1]$ be the normalized probabilistic output of $\mathcal{N}_m(\mathbf{x}_m)$ for category $c \in \{1, ..., C_m\}$, such that $\sum_{c=1}^{C_m} p_m(c|\mathbf{x}) = 1$. We convert these classification predictions into a set of text-embedding vectors by (1) sampling K_m categories (without replacement) from the probabilistic outputs for each modality m and then (2) embedding the names of the sampled categories via a matrix multiplication:

$$\mathbf{e}_m^k = \mathbf{W}_m^T \mathbf{c}_m^k. \tag{1}$$

where $\mathbf{W}_m \in \mathbb{R}^{C_m \times D}$ is a learned D-dimensional embedding of C_m category tokens and $\mathbf{c}_m^k$ is a one-hot vector that encodes the name of the k-th sampled category from modality m.

Note that the sampling process is necessary during training because directly selecting top predictions will drop the rich information in the predicted distributions and bias the training process [18]. In order to make the sampling differentiable, we leverage the Gumbel-Softmax trick [18] and a differentiable approximation of tokenization [8]. Specifically, we reparameterize the predicted probability distribution $\mathbf{p}_m \in \mathcal{R}^{1 \times C_m}$ by adding Gumbel noise $\mathbf{g}_m \in \mathcal{R}^{1 \times C_m}$ to it, where $\mathbf{g}_m = -\log(-\log(\mathbf{u}))$ with $\mathbf{u} \sim \text{Uniform}(0, 1)$. We then sample the top K_m categories from the reparameterized distribution $\tilde{\mathbf{p}}_m \in \mathcal{R}^{1 \times C_m}$ for each modality m.

With this re-parameterized distribution, selecting the top K_m categories is equivalent to sampling K_m categories from the original distribution. For detailed proof, we refer the reader to [22]. However, the process of selecting the top K_m categories is still not differentiable. To address this issue, we use a Straight-Through Estimator [18]. Specifically, during forward propagation, we sample top K_m categories as described above. Instead, during backward propagation we estimate the gradient for each category c as:

$$G \approx \nabla_{\mathbf{W}_m} \frac{\exp\left(\log p_m(c|\mathbf{x}) + \mathbf{g}_m(c)\right)}{\sum_{c'}^{|\mathcal{C}_m|} \exp\left(\log p_m(c'|\mathbf{x}) + \mathbf{g}_m(c')\right)}. \tag{2}$$

This leads to a unified formulation, which enables end-to-end learning of the entire system including the modality-specific classifiers. Furthermore, note that the embedding transformation $\mathbf{W}_m$ can be initialized using a pretrained language embedding space [40]. This simple procedure provides the advantage of converting all modalities into the same semantic language space, thus eliminating the need for designing complex cross-modal fusion blocks. Furthermore, we can seamlessly leverage powerful language encoders for our target task, which is highly beneficial.

3.2. Generative Encoder-Decoder

With the different modalities embedded in the same language space, we can directly use a text encoder to fuse the multimodal information. We collect the embedding vector $\mathbf{e}_t$ representing the task definition t together with the embeddings computed from the different modalities into a sequence of L vectors which we feed into the text encoder $\mathcal{F}_{En}$:

$$\mathbf{z} = \mathcal{F}_{En}(\mathbf{e}_t, \mathbf{e}_\text{S}, \mathbf{e}_1^1, ..., \mathbf{e}_1^{K_1}, \mathbf{e}_\text{S}, ..., \mathbf{e}_\text{S}, \mathbf{e}_M^1, ..., \mathbf{e}_M^{K_M}), \quad (3)$$

where $\mathbf{e}_\text{S}$ is the embedding of a special "separator" token, and $\mathbf{z} \in \mathbb{R}^{L \times d'}$ is a sequence of L vectors with a dimensionality d'. The features $\mathbf{z}$ produced by the text encoder capture task-specific information from multiple modalities.

Afterwards, we feed the new representation $\mathbf{z}$ to the decoder for text generation. Our decoder generates results in an auto-regressive manner, meaning that it uses previously decoded outputs as part of its input. Formally, we can write this as follows:

$$\hat{\mathbf{h}}_i = \mathcal{F}_{De}(\mathbf{z}, \tilde{\mathbf{h}}_1, ..., \tilde{\mathbf{h}}_{i-1}), \quad (4)$$

where $\hat{\mathbf{h}}_i \in \mathbb{R}^{T'}$ is the i-th decoded distribution over a dictionary of T' tokens and $\{\tilde{\mathbf{h}}_1, ..., \tilde{\mathbf{h}}_{i-1}\}$ are discrete history tokens. The decoding process will terminate when the "End-of-Sequence" token is generated.

3.3. Training

During training, we follow the common practice of teacher-forcing [49, 40], which means that we replace the decoding history with ground-truth tokens $\mathbf{h}_i$ in the corresponding positions:

$$\hat{\mathbf{h}}_i = \mathcal{F}_{De}(\mathbf{z}, \mathbf{h}_1, ..., \mathbf{h}_{i-1}), \quad (5)$$

Our entire system is then trained with a standard cross-entropy loss:

$$\mathcal{L} = \min_{\mathbf{w}} \frac{1}{n} \sum_i \text{Cross-Entropy}(\hat{\mathbf{h}}_i, \mathbf{h}_i), \quad (6)$$

where n is the number of valid tokens. Note, that this design supports generation of text with variable length. While here we show the objective for a single training sample, in practice we optimize over mini-batches of samples.

3.4. Inference

Most previous multimodal transformers [55, 29] rely on task-specific heads to tackle different tasks. Specifically, the heads designed for generative tasks typically differ substantially from those used in discriminative settings. However, our VX2TEXT seamlessly addresses both types of tasks without the need to change its architecture.

For generative tasks, e.g., captioning and video dialog, we follow previous works and use Beam Search [41, 25] (with beam width set to 5) or Greedy Decoding [28] to generate coherent sentences. Instead, for discriminative tasks, e.g., question answering on TVQA, the model is required to pick the most probable answer from a provided candidate set. In such cases, we include the entire set of candidate answers as additional input to the model (using separator tokens to mark them) and then evaluate each candidate output under the probability distribution defined by the autoregressive decoder. Finally, we select the highest-probability answer among the choices as the prediction. In this way, with a unified encoder-decoder structure, our model can handle both generative and discriminative tasks. In our experiments we demonstrate that the knowledge stored in the decoder helps our generative VX2TEXT outperform its discriminative counterpart as well as previous discriminative models (see Sections 4.4 and 4.5).

3.5. Implementation Details

We use R(2+1)D-34 [47, 15] trained on Kinetics [9] as our video backbone network, with the 400 action categories of Kinetics as the video vocabulary. We follow the video preprocessing procedure described in [15] to sample a clip of 32 frames. We sample $K_v = 12$ predicted categories from the pool to represent the action/events in videos.

As the audio backbone, we use CNN14 [21] trained on AudioSet [14] to recognize 527 acoustic events. Audio segments are sampled at 16,000 Hz from corresponding video clips. They are then processed to extract Log-mel spectrograms which are fed into the CNN. We use $K_a = 6$ predicted categories to represent the acoustic events in audio segments. We provide analyses on the hyperparameters K_v and K_a in the Appendix.

We note that our VX2TEXT is not constrained to use the tokens of the predefined categories for each modality. In the Appendix we present experiments where we map each modality into the the full vocabulary of tokens by using the predefined categories merely as an initialization. We show that this more general scheme yields equivalent results.

We use T5-base [40] as our text transformer including the text token embedding layer, the encoder and the decoder. We use pretrained weights provided in HuggingFace [50] for initialization of the text transformer. We note that, except for these initializations, we do not use any form of pretraining and that the optimization of the model is done on

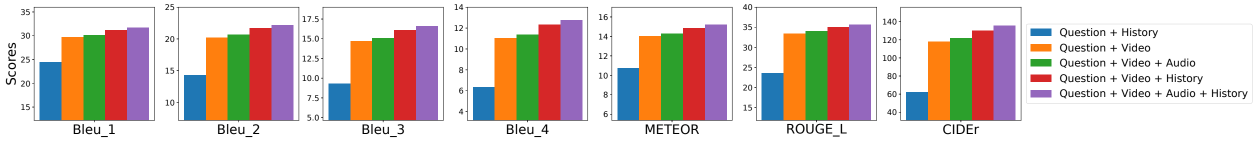


Figure 2. Impact of different combinations of multimodal inputs on the performance of VX2TEXT for the task of audio-visual scene-aware dialog on the AVSD validation set. (Best viewed in colors.) Each modality contributes to elevate the performance, especially the video input.

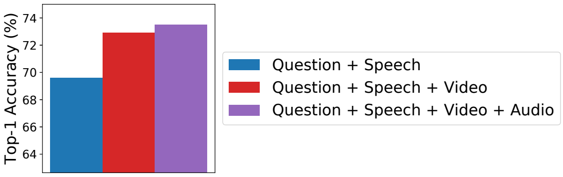


Figure 3. Video question answering performance of VX2TEXT on the TVQA validation set for different combinations of input modalities. (Best viewed in colors.)

each individual task using the given training set.

We use a batch size of 6 examples per GPU, and distribute the training over 32 NVIDIA V100 GPUs. We use Adam with a learning rate of 0.0001 to optimize our models. We train our models for 40 epochs, with the learning rate divided by 10 at the 20-th and 30-th epochs. Training on AVSD, TVQA, and TVC with our default settings takes about 12, 15, and 20 hours, respectively.

4. Experiments

In this section, we evaluate the effectiveness of VX2TEXT on three distinct tasks: (1) video-based question answering, (2) audio-visual scene-aware dialog, and (3) video captioning. We use three benchmark datasets: TVQA, AVSD, and TVC for these three tasks, respectively.

4.1. Datasets and Evaluation Metrics

Audio-Visual Scene-Aware Dialog. AVSD [2] is a benchmark of human dialogs describing videos in the Charades dataset [43]. The dialogs are in the form of 10 question-answer (QA) pairs per video. The questions are formulated by a human subject who has not observed the video. The questions aim at collecting as much information as possible about the content of the video. A person who has seen the video provides detailed answers to the questions through the dialog. Algorithms are evaluated on this benchmark by their ability to answer the questions in textual form. As in prior work [41], we adopt the following evaluation metrics: BLEU-{1,2,3,4} [37], CIDEr [48], METEOR [7], and ROUGE-L [32].

Video Question Answering. TVQA [26] is a dataset of video clips collected from 6 TV series. Given a video clip and its corresponding speech, the goal of this task is to answer a multiple-choice question about the clip. Each video clip has 7 questions, with 5 candidate answers per question. In total, the dataset consists of 152,500 QA pairs from 21,800 clips. The speech data comes in the form of manually annotated transcripts.

Video Captioning. TVC [28] is a recently introduced benchmark for video captioning. The TVC dataset includes the same set of videos as TVQA, but the videos are segmented into clips in a different way. We follow the protocol introduced in previous work [28] and include the speech consisting of manual transcripts as input to our model. We adopt the following evaluation metrics: BLEU-{1,2,3,4} [37], CIDEr[48], METEOR [7], and ROUGE-L [32].

4.2. Assessing the Importance of Each Modality

We begin by studying the effect of individual modalities on video-based text generation performance. We do so by training and testing our model with different combinations of inputs. Results are shown in Figure 2 for the AVSD dataset, and in Figure 3 for the TVQA dataset. Based on these results, we observe that each modality provides a performance gain for both tasks. This is especially noticeable for the AVSD benchmark, which was specifically designed for multimodal understanding. Furthermore, note that the addition of the video modality yields a very significant gain under all metrics on AVSD compared to the version of our model relying only on textual input (question and history). This trend also holds on the TVQA dataset. Finally, we also observe that leveraging the history of previous QA pairs is highly beneficial on AVSD. This suggests that our model successfully incorporates information from previous QA pairs in the dialog.

Although previous studies [26, 41] have shown that these benchmarks are somewhat biased to input text, our model improves from text-only settings significantly, e.g., by **14.4%** on AVSD and by **3.5%** on TVQA. The larger gain compared to prior works [26, 41] reveals the strength of our proposed video encoding and fusion strategy.

4.3. The Effect of Differentiable Tokenization

In this section, we demonstrate the usefulness of our proposed Differentiable Tokenization scheme. For this

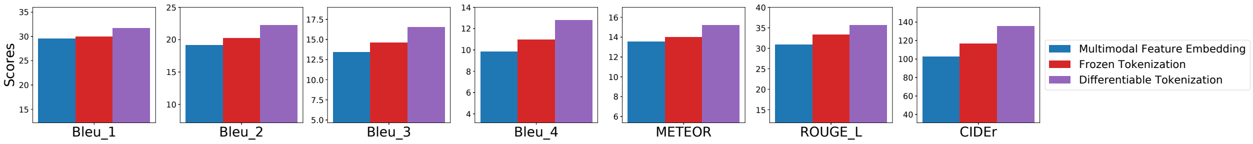


Figure 4. Comparing performance obtained with our Differentiable Tokenization vs the baselines of Multimodal Feature Embedding and Frozen Tokenization on the AVSD validation set. (Best viewed in colors.) Differentiable Tokenization enables end-to-end training with respect to the end objective and yields the best performance.

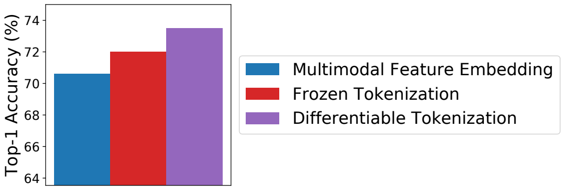


Figure 5. Studying the effect of different modality fusion mechanisms on the QA performance of the system on the TVQA validation set. (Best viewed in colors.) Differentiable Tokenization outperforms the other schemes by large margins.

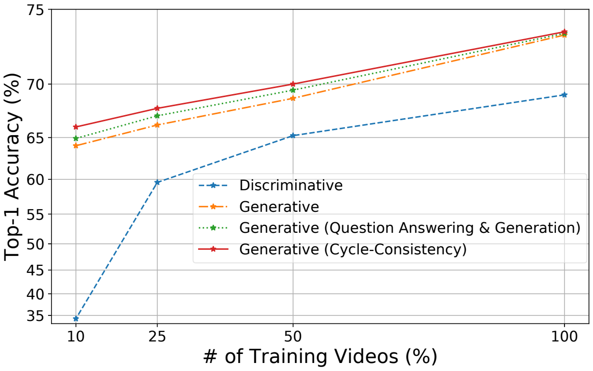


Figure 6. Comparison between a Discriminative variant of our VX2TEXT and the default Generative version on TVQA. The generative version achieves much higher accuracy for all training set sizes. Furthermore, the generative formulation enables multi-task learning (see "Question Answering & Generation" and "Cycle-consistency") with the *same* model. This yields further improvements in accuracy, especially for small training sets.

purpose, we consider and test two comparative baselines. The first, named Multimodal Feature Embedding, uses a modality-specific fully-connected layer with Layer Normalization [6] to map the continuous predictions of the audio and video classifiers into the language embedding space. This scheme is similar to the strategy implemented by the input embedding modules in HERO [29] and it provides an alternative way to enable end-to-end training.

For the second baseline, we replace our Differentiable Tokenization with Frozen Tokenization, which means that only the text transformer is trained with respect to the target task, while the modality-specific networks are frozen to their pretrained configurations. The results are shown in Figures 4 for AVSD and in Figure 5 for TVQA, using all available input modalities for both tasks. It can be observed that Frozen Tokenization achieves better performance than the Multimodal Feature Embedding. This by itself already provides evidence of the benefit obtained by mapping all modalities into the language space using the top predictions of the modality-specific classifiers. However, it can be noticed that Differentiable Tokenization boosts further the performance on both tasks by jointly optimizing the entire model end-to-end.

4.4. The Benefit of a Generative Model

To show the benefits of our unified generative formulation, we present a comparison involving four models trained and evaluated on TVQA. The first model is our default VX2TEXT model, denoted here as Generative. The second model is a discriminative version of our system obtained by removing the decoder and by attaching a classification head to the pooled embedding obtained from the encoder. This variant is trained end-to-end to predict a distribution over the five candidate answers. It is similar to the approach taken in HERO [29], except that it uses our Differentiable Tokenization as modality fusion mechanism. As a reference, we found that our Discriminative baseline achieves performance comparable with that of HERO (without pretraining) on TVQA.

Furthermore, to show the flexibility of our generative formulation, we include two additional variants of VX2TEXT using multiple generative training objectives. "Generative (Question Answering & Generation)" has two training objectives: one is for video question answering and the other is for video question generation. When generating questions, our model takes $Question$ as the task token t and the ground-truth answer as part of the input. In such mode the system is asked to predict in a generative manner the ground-truth question from the ground-truth answer.

In "Generative (Cycle-consistency)", our model performs the following steps: 1) generates answer A' given the ground-truth question Q; 2) produces question Q'' based on A'; 3) outputs answer A'' based on Q''. The final objective is a linear combination of the Question consistency

Models	Use Caption?	CIDERr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR
MA-VDS [16]	No	0.727	0.078	0.109	0.161	0.256	0.277	0.113
Simple [41]	No	0.905	0.095	0.130	0.183	0.279	0.303	0.122
VX2TEXT (Ours)	No	**1.357**	**0.127**	**0.166**	**0.222**	**0.317**	**0.356**	**0.152**
MTN [25]	Yes	1.249	0.128	0.173	0.241	0.357	0.355	0.162
MTN-TMT [30]	Yes	1.357	0.142	-	-	-	0.371	0.171
VX2TEXT (Ours)	Yes	**1.605**	**0.154**	**0.197**	**0.260**	**0.361**	**0.393**	**0.178**

Table 1. Comparison to the state-of-the-art on the AVSD test set with and without caption as input. Our model achieves the best results under both settings.

Models	# Samples for Multimodal Pretext	Val	Test
HERO [29]	7.6M	74.8	73.6
TVQA [26]	0	67.7	68.5
STAGE [27]	0	70.5	70.2
HERO [29]	0	70.7	70.3
MSAN [20]	0	71.6	71.1
BERT QA [52]	0	72.4	72.7
VX2TEXT (Ours)	0	**74.9**	**75.0**

Table 2. Comparison to the state-of-the-art for the task of Video Question Answering on both the validation set and the test set of TVQA. On the test set, VX2TEXT achieves even better performance than the version of HERO that leverages 7.6M additional multimodal samples for pretraining. Top-1 Accuracy (%) is reported.

$|Q'' - Q|$, the Answer consistency $|A'' - A|$ as well as the Question Answering and Question Generation losses. Such a multi-loss objective was originally proposed by Shah et al. [42] for the case of image-based QA. For details of these two baselines, please refer to our Appendix.

Figure 6 shows the performance of these four models as we vary the number of QA pairs used for training. Our VX2TEXT model trained in a generative fashion significantly outperforms its discriminative counterpart for all training set sizes, but especially so when data is dramatically reduced. For example, the accuracy gap between Generative and Discriminative is 29.9% (64.1% vs 34.2%) when using 10% of the training data. We believe that this large performance difference comes from the beneficial commonsense knowledge stored in the text decoder.

Moreover, our generative formulation allows VX2TEXT to be trained with respect to multiple tasks without the need to change the architecture or add network heads. As shown in Figure 6, this translates in further performance improvements. For example, with the help of Cycle-consistency, our VX2TEXT achieves an accuracy of 66.1% (vs the 64.1% of Generative) when using 10% of the data. Our VX2TEXT trained with Cycle-consistency using only 50% of the training samples outperforms the Discriminative model trained on the full training set (100% of the samples).

4.5. Comparison With the State-of-the-Art

We now compare our single architecture separately trained on the three benchmarks to the state-of-the-art.

AVSD. Our comparative results on this benchmark are shown in Table 1. Our VX2TEXT significantly improves over existing methods both with and without text caption as part of the inputs. Note that the state-of-the-art MTN system [25] uses complex cross-modal attentional modules to fuse the information from different modalities. MTN-TMT [30] leverages complex auxiliary losses to align the embedding spaces of MTN. However, even without text caption, which is a very strong information source, our VX2TEXT achieves already better performance than MTN. When adding text caption to the input, the performance of our VX2TEXT is further boosted and it significantly surpasses that of MTN-TMT. This further demonstrates the effectiveness of our simple scheme for modality integration.

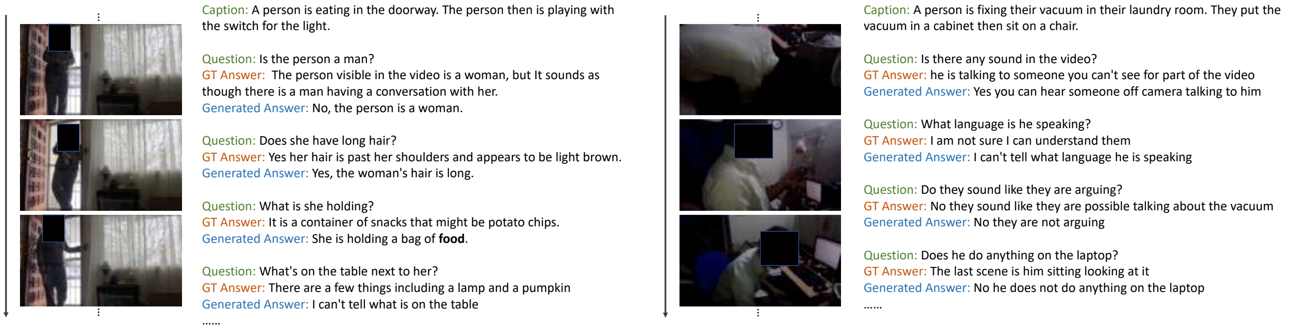
TVQA. Since many methods on TVQA use object/frame-level features, for a fair comparison, we include detected object categories [26] as an extra modality of input for VX2TEXT in this evaluation. Due to the complexity of training object detectors, here we use Frozen Tokenization and leave the application of Differentiable Tokenization for future work.

Table 2 shows that on TVQA our VX2TEXT significantly outperforms all previous methods on both the validation set and the test set when training is done without additional multimodal pretext training data. On the test set, our VX2TEXT yields an improvement of 1.4% compared to the previous state-of-the-art, represented by the HERO system which adopts an expensive multimodal pretext training on 7.6M additional samples. As reported in [29], this pretraining takes about 3 weeks. When both models are trained without multimodal pretext, our VX2TEXT outperforms HERO by 4.7%.

TVC. Table 8 shows that on the captioning task of TVC our VX2TEXT significantly outperforms the state-of-the-art MMT [28] system. Without pretraining, HERO achieves performance comparable to that of MMT and inferior to ours. With multimodal pretraining on additional 7.6M samples (again requiring 3 weeks), HERO does only slightly better than our model. Our VX2TEXT also shows good generalization on the test set. Note that, as done on TVQA, even here we include object detection predictions as an input modality for our model since the methods considered in this comparison all have access to frame-level features.

Models	# Samples for Multimodal Pretext	Validation				Test			
		CIDERr	BLEU-4	ROUGE-L	METEOR	CIDERr	BLEU-4	ROUGE-L	METEOR
HERO [29]	7.6M	0.505	0.123	0.341	0.175	0.500	0.124	0.342	0.176
MMT [28]	0	0.444	0.105	0.324	0.166	0.454	0.109	0.328	0.169
HERO [29]	0	0.436	0.107	0.327	0.164	0.437	0.109	0.326	0.165
VX2TEXT (Ours)	0	**0.482**	**0.116**	**0.328**	**0.172**	**0.483**	**0.119**	**0.331**	**0.174**

Table 3. Video captioning performance of VX2TEXT on both the validation set and the test set of TVC. Our model achieves the best performance among the methods that do not make use of additional samples for multimodal pretraining.



Caption: A person is eating in the doorway. The person then is playing with the switch for the light.

Question: Is the person a man?
GT Answer: The person visible in the video is a woman, but It sounds as though there is a man having a conversation with her.
Generated Answer: No, the person is a woman.

Question: Does she have long hair?
GT Answer: Yes her hair is past her shoulders and appears to be light brown.
Generated Answer: Yes, the woman's hair is long.

Question: What is she holding?
GT Answer: It is a container of snacks that might be potato chips.
Generated Answer: She is holding a bag of **food**.

Question: What's on the table next to her?
GT Answer: There are a few things including a lamp and a pumpkin
Generated Answer: I can't tell what is on the table
......

Caption: A person is fixing their vacuum in their laundry room. They put the vacuum in a cabinet then sit on a chair.

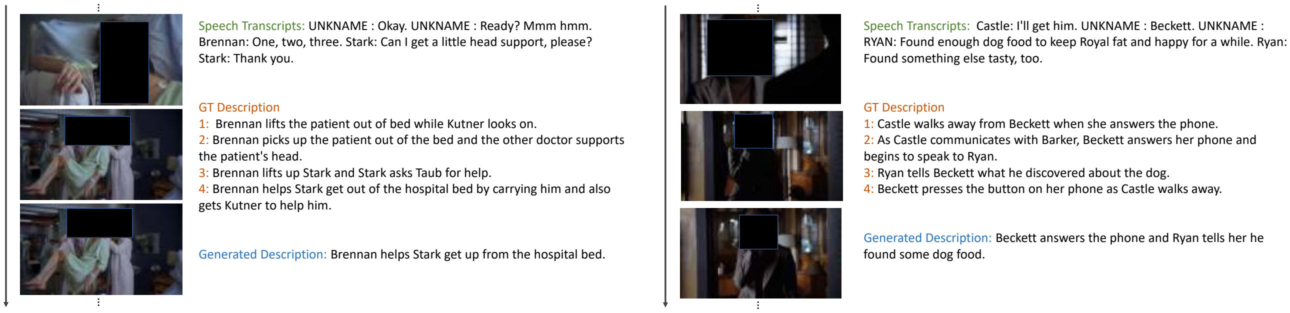
Question: Is there any sound in the video?
GT Answer: he is talking to someone you can't see for part of the video
Generated Answer: Yes you can hear someone off camera talking to him

Question: What language is he speaking?
GT Answer: I am not sure I can understand them
Generated Answer: I can't tell what language he is speaking

Question: Do they sound like they are arguing?
GT Answer: No they sound like they are possible talking about the vacuum
Generated Answer: No they are not arguing

Question: Does he do anything on the laptop?
GT Answer: The last scene is him sitting looking at it
Generated Answer: No he does not do anything on the laptop
......

Figure 7. Examples of generated answers for audio visual scene-aware dialog on the AVSD validation set. Our VX2TEXT successfully responds in natural language given the multimodal inputs. Faces in the frames are artificially masked for privacy reasons.



Speech Transcripts: UNKNAME : Okay. UNKNAME : Ready? Mmm hmm. Brennan: One, two, three. Stark: Can I get a little head support, please? Stark: Thank you.

GT Description
1: Brennan lifts the patient out of bed while Kutner looks on.
2: Brennan picks up the patient out of the bed and the other doctor supports the patient's head.
3: Brennan lifts up Stark and Stark asks Taub for help.
4: Brennan helps Stark get out of the hospital bed by carrying him and also gets Kutner to help him.

Generated Description: Brennan helps Stark get up from the hospital bed.

Speech Transcripts: Castle: I'll get him. UNKNAME : Beckett. UNKNAME : RYAN: Found enough dog food to keep Royal fat and happy for a while. Ryan: Found something else tasty, too.

GT Description
1: Castle walks away from Beckett when she answers the phone.
2: As Castle communicates with Barker, Beckett answers her phone and begins to speak to Ryan.
3: Ryan tells Beckett what he discovered about the dog.
4: Beckett presses the button on her phone as Castle walks away.

Generated Description: Beckett answers the phone and Ryan tells her he found some dog food.

Figure 8. Examples of textual descriptions generated by VX2TEXT for video captioning on the TVC validation set. Our VX2TEXT generates informative descriptions from multimodal inputs. Faces in the frames are artificially masked for privacy reasons.

4.6. Qualitative Results

As shown in Figures 7 and 8, our VX2TEXT generates realistic natural text for both audio-visual scene-aware dialog and video captioning. It is very encouraging that although our model takes some text inputs, e.g., dialog histories or speech transcripts, the generated text does include information from other modalities. For example, as Figure 8 shows, our model successfully recognizes the actions, e.g., helping to get up or answering the phone, and even grounds the characters correctly. In the Appendix, we further investigate and show the semantics of the predicted tokens.

5. Conclusions

In this work we have presented a simple unified framework to address the problem of text generation from video with additional modalities. Our approach hinges on the idea of mapping all modalities into a semantic language space in order to enable the direct application of transformer networks, which have been shown to be highly effective at modeling language problems. We have introduced a mechanism of differentiable tokenization to convert the continuous outputs of modality-specific classifiers into the language space. This renders our entire model trainable end-to-end. Our framework applied to a single architecture outperforms the state-of-the-art on three different video-based text-generation tasks.

Acknowledgments

Thanks to Bruno Korbar, Rohit Girdhar and Fabio Petroni for their work in an early version of this project. This research was done while Xudong Lin was an intern at Facebook AI.

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 Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019.

[2] Huda Alamri, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Jue Wang, Irfan Essa, Dhruv Batra, Devi Parikh, Anoop Cherian, Tim K Marks, et al. Audio visual scene-aware dialog (avsd) challenge at dstc7. *arXiv preprint arXiv:1806.00525*, 2018.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[7] 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.

[8] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[11] 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.

[12] 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.

[13] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304, 2015.

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

[15] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019.

[16] 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.

[17] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002, 2020.

[18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[19] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766, 2017.

[20] Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. Modality shifting attention network for multi-modal video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10115, 2020.

[21] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[22] Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. *arXiv preprint arXiv:1903.06059*, 2019.

[23] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169, 2018.

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[25] Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, 2019.

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

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

[28] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. *arXiv preprint arXiv:2001.09099*, 2020.

[29] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[30] Wubo Li, Dongwei Jiang, Wei Zou, and Xiangang Li. Tmt: A transformer-based modal translator for improving multimodal sequence representations in audio visual scene-aware dialog. *arXiv preprint arXiv:2010.10839*, 2020.

[31] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[32] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[33] 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.

[34] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016.

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

[36] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2017.

[37] 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.

[38] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[39] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

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

[42] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6649–6658, 2019.

[43] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

[44] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning, 2019.

[45] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[46] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[47] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[48] 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.

[49] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[50] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910, 2019.

[51] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 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, 2016.

[52] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. Bert representations for video question answering. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1556–1565, 2020.

[53] Xuwang Yin and Vicente Ordonez. Obj2text: Generating visually descriptive language from object layouts. *arXiv preprint arXiv:1707.07102*, 2017.

[54] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278*, 2015.

[55] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pretraining for image captioning and vqa. *arXiv*, pages arXiv–1909, 2019.

[56] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. *arXiv preprint arXiv:1703.09788*, 2017.

[57] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[58] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.