Causal Understanding For Video Question Answering

Adithya Sampath [*1] **Bhanu Prakash** [*1] **Tanmay Kulkarni** [*1] **Swarnashree Mysore Sathyendra** [*1]

1. Introduction

In this report, we aim to perform a detailed analysis of the NExT dataset (Xiao et al., 2021) to uncover any hidden biases, and to estimate the hardness and possibility of modeling the task of VideoQA using deep learning approaches, in particular the need for multimodal modeling. This report is thus divided into: Idea specific analysis[2] that includes human-in-the-box experiments[2.1] and qualitative analysis[2.2], Modality specific analysis[3], where we look at unimodal text analysis[3.1] and unimodal video analysis[3.2] to understand how much each modality influences the downstream task and thus eliciting the need for multimodal models for VideoQA.

2. Idea-Specific Analysis

To reiterate, our proposed ideas include finding spurious biases in the dataset, modeling multiframe reasoning models owing to the need for answering *causal* and *temporal* questions, and the ability to find a good counterfactual for the irrelevant frames for answering questions. Towards this end, we performed some human-in-the-box experiments and also some qualitative analysis on the obtained annotations from the human experiments. We conclude this section by providing a quantitative correlation analysis to uncover the possibility of some hidden biases in the data annotation procedure developed by (Xiao et al., 2021).

2.1. Human In the Box

Our human in the box experiments involve annotations from all the team members [1]. All the experiments involve random sampling of 15 videos from the dataset and annotations to the following framed tasks from all the four annotators. We have framed the following 3 tasks to understand the need for multiframe reasoning and the need for multimodal models.

1. **T1:** Can you answer the question looking at the question alone and the answer choices?
2. **T2:** Can you leverage only the first, middle and last frames of the video to answer a question?
3. **T3:** Would you require more than one frame to correctly estimate the answer (need for more info)?

[1]Detailed annotations are available here

Task	Accuracy	Fleiss Kappa
T1	0.467	0.447
T2	0.867	0.611
T3	-	0.154

Table 1. Accuracy and inter-annotator agreement for the human annotation tasks

Table 1 reports the results for the aforementioned tasks. We will now discuss the motivation and the interpretation for these different tasks. The performance of task **T1** indicates *text alone is sometimes sufficient to answer the questions*. Often, we as humans, used external knowledge and plausibility from language to answer the questions. This also indicates that some bias could be possible from the modeling which could use the *text features alone could bias the answers*. However, the low accuracy indicates that still a good number of questions require additional information to accurately predict the answer. Including just three frames of a video with the text modality in task **T2** increases the accuracy to 86.7%. This indicates the potential information gain from a multimodal approach at a human judgement level, motivating us to look at *multimodal modeling approaches for the task at hand*. We designed the task **T3** to understand more about the need for multiframe information for reasoning the answer to the question (note that accuracy is not a metric and hence is omitted from table). Out of the 15 random questions with video information from the first, middle and last frames, the annotators reported that [8, 8, 9, 10] number of questions needed more information for the four different annotators indicating *at least half of the questions requiring for even more frames to predict the correct answer*.

We also use Fleiss Kappa(Fleiss, 1971) to analyse the agreement between the annotators. Table 1 shows considerable agreement between annotators in tasks **T1** and **T2**. However, in the absence of video information in task **T1**, *even humans find it hard to answer the questions and agree upon the answer*. Task **T3** has a surprisingly low agreement score. However through a conversation amongst the annotators, we have concluded that different annotators needed more information for answering different questions, which in turn would mean that on an average more information would be needed for each data point. Each question had at least

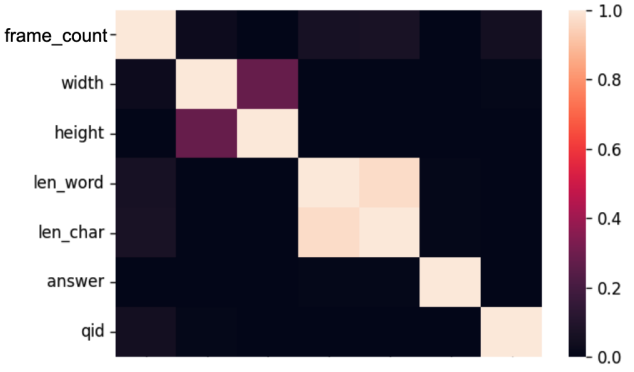


Figure 1. Correlation values between different attributes of the questions and videos and the final answers in the mcq type questions.

one annotator reporting that more information is needed and the total count distribution per question had a *median of 3 annotators*. Overall these results call for a multiframe and multimodal reasoning model, although some bias is possible from text alone modality.

2.2. Qualitative Analysis

Upon observing a high disagreement in **T3** and also to some extent in **T1** and **T2**, we performed a qualitative analysis that could aid us while developing the final multimodal model for VideoQA. A couple of example questions where the annotators disagreed on the answers greatly are:

1. *Why are the people in formal wear in the room?*
2. *How does the baby interact with the blanket?*
3. *why does the boy in black suck his thumb?*

The first question is a data point from the task *T1* and the annotators have observed that it is hard to solve with only text data and confused between *work* and *party* options, as *both options are equally plausible in absence of video*. In case of the task **T2**, questions like the second one from the list above need more information than simply three frames to *understand the interaction of the subject with the object over time*. While the text plausibility removes ambiguity amongst the answer choices to some extent, we need multiframe reasoning to answer this question rather than a single frame. In task *T3*, questions like the third one from the above list led to *some commonsense knowledge from the humans* to decide upon whether more information is needed or needed to answer the question. Commonsense knowledge to straightforward answer choice *habit* for the above question as there wasn't any harm to the hand in the observed three frames.

Overall, the annotators observed a relatively high success rate by exploiting the *grammatical similarities between the question and the expected answer*(e.g. a question on *'why'* more likely expects an answer that begins with the infinitive *to* rather than the gerund *-ing* form). However, they faced confusion amongst the answer choices when the annotations provided by the example in the dataset do not follow grammatical rules. In addition, quite a few answers can be decided on the basis of common sense reasoning. For instance, *why is the bird pecking the ground; ans - To find food*. The annotators also struggle when there seems to be no real context match between the question and the answer, even after proving video information.

These examples showcase some of the imperfections and hardness of the dataset. However, a common narrative is that when there are questions that involve items/subjects/objects that are not expected to change across video frame of the same video(eg: people in suits, unlikely to change in the course of the same video), annotators answer better, as compared to videos that have lot of actions (since they can keep changing across video) calling for a complex multimodal multiframe reasoning model.

2.3. Hidden Meta Biases

One of our research ideas include eliciting any kind of hidden dataset biases that are induced in the dataset during the annotation procedure. Towards this end, we have performed some correlation analysis and also trained a very simple model for predicting answers. Along with the dataset, we have some additional meta data that is available in the repository. We use the meta attributes of the questions and videos to perform a multi-class and linear correlation analysis which is reported in the Figure 1(*qid* corresponds to the question type). We observe correlation only between the expected attributes i.e *height* and *width* of the video, *question length* in terms of the number of *words* and *characters*. We don't observe any positive or negative correlation between any of the meta attributes of questions and videos to the answer indicating that hidden confounding factors are absent from these attributes point of view.

To observe if there are any anomalous correlations between a question and answer choice, we train a simple text classification model with question text as input and answer choice as predicted label, performance shown in Table 3. In this setting we simply provide the text of the question without answer text and perform a 5-class classification (there are 5 answer choices per question). We observe a very low performance of $\sim$20% which is almost equivalent to a random model for a 5-class classification setup. Hence, we conclude that there are no hidden spurious correlations between the questions and answer labels.

3. Modality-Specific Analysis

In this section we try to analyse the individual modalities of information which are the language and vision for the NExT-QA dataset.

3.1. Text Modality

For the text modality, we perform analysis using both multi-modal fine-tuned and general language based understanding of the text modality.

3.1.1. With GloVe and BERT embeddings of NExT-QA

We perform clustering based and visualization based analysis on the GloVe and BERT embeddings provided in the NExT-QA dataset. Since both these embeddings are provided at a word-level, each question/answer is converted to a corresponding embedding by averaging the word embeddings of that question or answer. As a pre-processing step, we perform PCA on the above features, to find principal components that preserve at least 90% of the variance. We then perform Mini-batch K-Means on the principal components of the video features to perform clustering on the data This analysis is performed on question-answer pair ($< q, a >$ pair) with a <pad> token in-between with answer being the right choice in the MCQ type questions, since the (Li et al., 2022) baseline paper uses the $< q, a >$ pair for training.

Fig.6 shows the visualizations from running K-means algorithm on the glove embeddings on the training dataset (34132 question-answer pairs). Since we know the number of question types, we attempt K-means with $K = 8$ (number of question types shown in Table.2). Figure 7a represents the clusters identified by the K-means algorithm using the fine-tuned BERT embeddings provided with the dataset. Figures 6 and 7 in the Appendix A.1 shows the plots for Glove embeddings also using question-types as the cluster labels. We have observed that descriptive based question types (DC, DL, DO) are most discriminant i.e. even an unsupervised k-means algorithm is able to identify them. However, the Causal and Temporal type question (CW, CH, TP, TN, TC) are not easily distinguishable. Even the t-SNE plots (refer Appendix A.2) for the $< q, a >$ pairs are in agreement with these observations. For BERT-based embeddings, Fig.7, the descriptive type questions clusters are distinguishable. In addition, even temporal type questions are more distinguishable than the GloVe embeddings case. This makes sense as BERT embeddings are contexualized and it is possible that some form of temporality is introduced into these embeddings as compared to GloVe embeddings.

Since our project focuses on better causal and reasoning modelling, this is imminent good news as this shows that the causal/temporal questions do not have any obvious patterns that the model could rely on to take shortcut methods for modeling. Next section discusses about the feasibility and model accuracies of using only the text embeddings to answer the questions.

Question Type	Question Sub-type
Causal	How(CH), Why(CW)
Temporal	Previous(TP), Next(TH), Current(TC)
Descriptive	Count(DC), Location(DL), Other(DO)

Table 2. Question types in NextQA dataset

Figure 2. K-means assigned cluster ID with $< q, a >$ BERT Embeddings.

3.1.2. With Pre-trained and Fine-tuned BERT

Pre-trained BERT In this section we analyze the language modality of the dataset at even more depth. First we try to analyze the questions alone without looking at the answer choices. For this, we simply extract question level embeddings from the BERT-large model and try to visualize them. To understand the questions from the general language point of view, unlike the previous analysis, we don't use the multimodal fine-tuned question embeddings. Instead, we use the publicly available BERT large model[2] to generate the text representations. In this setting, we choose the uncased model as the available questions and answers

[2]https://huggingface.co/bert-large-uncased

Model	Training mode	accuracy
BERT-large	no answers	19.35%
BERT-large	answer choice	46.11%

Table 3. Performance of language modality alone BERT models. No answer mode indicates just providing the question text and performing a 5-class classification.

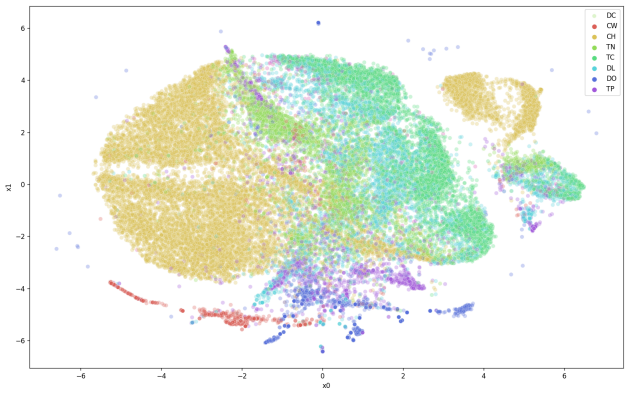


Figure 3. t-SNE analysis of the embeddings of the questions using pre-trained BERT-large model.

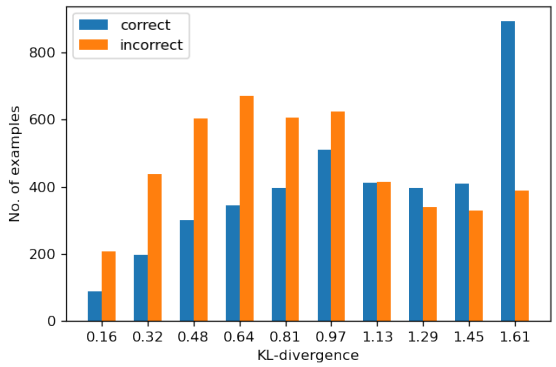


Figure 4. KL divergence between model assigned probabilities for the choices and complete random model i.e 0.2 probability for each choice.

are already preprocessed to lower case English alphabet by the authors of the NExT-QA dataset. We take the [CLS] token representation of the last layer and perform a 2d-tSNE analysis on the resulting embeddings. Figure 3 shows the visualization of the obtained clusters. Even using general language embeddings which are not fine tuned for video question answering task, we can visibly see that the questions are distinguishable by their type. Surprisingly, the question types that belong to the same higher level class (descriptive, temporal, and causal) belong to separable clusters in the visualization. This could be either due to the 2d projection where we have lost information about other dimensions of variation or some inherent language specific variations among the sub question class types. We intend to utilize this observation for our future modeling purposes.

Fine-tuned BERT In the next part, we utilize only the text information to predict the answers for the questions. In this setup, similar to the standard MCQ style question answering models, we concatenate each question with all its possible choices of answer and perform a n-class classification. The result for this experiment corresponds to the second row of the results reported in Table 3. As we can observe, we obtain a very good performance of ∼46% using the questions and answer choices alone. The implementation for this setup is based on the Hugging Face[3] implementation for the SWAG dataset. The intention behind performing this experiment is to remove the biases in the multimodal finetuned embeddings which are provided by the original authors. This setup completely relies on the text information unlike the previous subsection. This experiment also supplements our human analysis where we looked at the text alone and performed the classification. The performance from this model is in fact very close to the human performance. Table 4 reports some of the best and worst performing examples from this model. In case of incorrect predictions, the model answer also seems to be biased based on general knowledge or the similar words between questions and the answer choices. To further analyze the confidence of the model in these predictions, we have conducted the KL-divergence analysis. Here, we compare the model probability scores with the complete random model (assigning 0.2 probability to all choices). The histogram of the KL-divergence values for the correct and incorrectly predicted examples is shown in Figure 4. As observed from the figure, the correct predictions have a higher deviation from the random model i.e more confidence in the answer choices compared to the incorrect ones. In incorrect cases, it looks like the model isn't able to distinguish much between the answer choices, i.e calling for more information from the video modality.

3.2. Visual Modality

We perform unsupervised clustering and visualization on the video features provided in the NExT-QA dataset. The features are extracted as part of HCRN (Le et al., 2020) VQA model using using Resnet101 (He et al., 2016) and ResNeXt101 (Xie et al., 2017) to capture appearance and motion features. The features are of 16 sample frames in the video with shape (N, 16, 4096) - where N is the number of videos, and 4096 dimension is the concatenation of Resnet and ResNeXt features. Since we want a condensed feature representation of the video, we take an average of all the 16 frames (i.e. along axis=1) to get a (N, 4096) feature vector. This condensed representation is required for clustering since algorithms like K-Means require 2D features. Another approach we experimented with was to cluster using only one frame number of all videos. Here, N = 3870, which is the number of videos in the train set. The number of QA pairs in the NeXT train dataset is 34132, hence, there are multiple QA pairs per video. Each video has one question of each type, and sometimes, more than one question of each type.

[3]https://huggingface.co/docs/transformers/tasks/multiple_choice

Question	Type	Answer	Model Answer
why are there flashes seen throughout the video	CW	camera flash	camera flash
how are the children prevented from staining their clothes while eating	CH	wear bibs	wear bibs
why are there flashes seen sometimes in the video	CW	camera taking pictures	camera taking pictures
why are there people standing on the roads that are meant for cars near the end of the video	CW	talking	waiting to cross
how did the people make sure they can see clearly in the dark	DO	shine light	wear spectacles
what did the lady in pink do after picking up the brush at first	TN	move it on the egg slowly	pick up more brushes

Table 4. Qualitative analysis of the best (top 3) and worst (bottom 3) performing test examples.

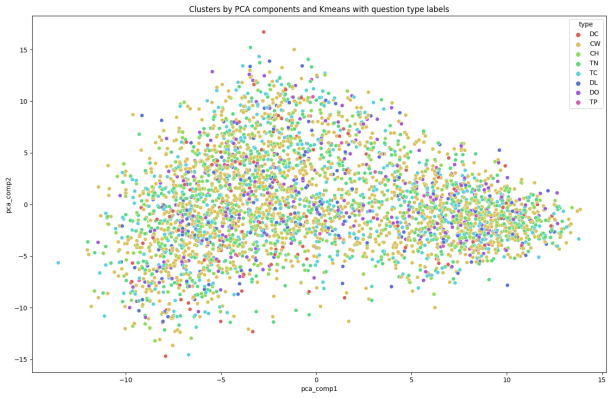


Figure 5. Question type labels based k-means clustering for first 3870 QA pairs using Resnet and ResNeXt features.

As a pre-processing step, we perform PCA on the above video features, to find principal components that preserve at least 90% of the variance. We then perform Mini-batch K-Means on the principal components of the video features to perform clustering on the data. We use nclusters=8 since that's the number of question types. The two experiments we performed are as follows: (1) For the 3870 video features assign cluster labels for only the first 3870 QA pairs in the train set; (2) do the same for the next 3870 QA pairs. We have observed that the video feature clusters are invariant to the question type label (refer to figures 9a and 9c in Appendix A.3). From Figure 9b (and also Figure 9d in Appendix A.3) we can observe that for a given video with with different question label types, there is no specific clustering of videos that entail a causal question / casual scene. Thus, thus our VideoQA can't leverage only video features to identify patterns in questions and take shortcuts to find answers.

Since the HCRN model was fine-tuned on the NExT-QA data, we wanted to ensure that our results aren't biased. To ensure that our hypothesis hold true, we perform a similar experiment with S3D and ViT features extracted from around 600 videos from our train set. For the S3D features we use the S3D Text-Video model trained on the HowTo100M dataset using MIL-NCE loss (Miech et al., 2019; 2020; Hara et al., 2018). Similary, for the ViT-based features, we use the VideoMAE model, essentially a ViT with a masked AutoEncoder architecture, trained on the Kinetics-400 dataset (Tong et al., 2022; Dosovitskiy et al., 2020; Kay et al., 2017). Even with ViT and S3D features, we observed that our hypothesis holds true, i.e., video features can't be clustered based on the question-type labels (Figures reported in Appendix A.4). Upon performing tSNE on the aggregated features, we again observed that the video features aren't directly correlated to question-type labels. The clustering images are reported in Appendix A.5

4. Re-Evaluation of Proposed Ideas

Based on the analysis that we have presented so far, we conclude the following regarding the proposed ideas.

1. Commonsense reasoning to some extent help answering the questions with text only information or single frame information. However, we need more contextual information from the causal and temporal reasoning of the videos to accurately answer the questions and achieve a better performance through multimodal models.
2. The videos did not really require any higher level reasoning. In fact most videos that we sampled could be solved by using a few frames. However, there were some videos that did require some amount of multi-frame reasoning.
3. Biases do present in the dataset from the language modality point of view. A decent accuracy with text alone representations in both human and model performance calls for a careful modeling of the final multimodal model, to avoid any spurious biases from text and properly leverage information from the multimodal models.
4. The videos follow a skewed distribution similar to that of VidOR(Shang et al., 2019; Thomee et al., 2016). This means that most of the scenes have human beings interacting with one-another. This makes it possible for finding better modeling based approaches to find hard-negatives from similar scenes i.e., the counterfactual extraction.

5. We have also observed that the dataset do not need a very strong causal reasoning model to answer even the causal questions. Several of the causal questions from the dataset can be answered using multiple frame analysis rather than as a chain of events in the video.

Github Repository

https://github.com/11777-MMML/11777-videoQA

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Hara, K., Kataoka, H., and Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6546–6555, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Le, T. M., Le, V., Venkatesh, S., and Tran, T. Hierarchical conditional relation networks for video question answering. *arXiv preprint arXiv:2002.10698*, 2020.

Li, Y., Wang, X., Xiao, J., Ji, W., and Chua, T.-S. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2928–2937, 2022.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., and Zisserman, A. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020.

Shang, X., Di, D., Xiao, J., Cao, Y., Yang, X., and Chua, T.-S. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 279–287. ACM, 2019.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

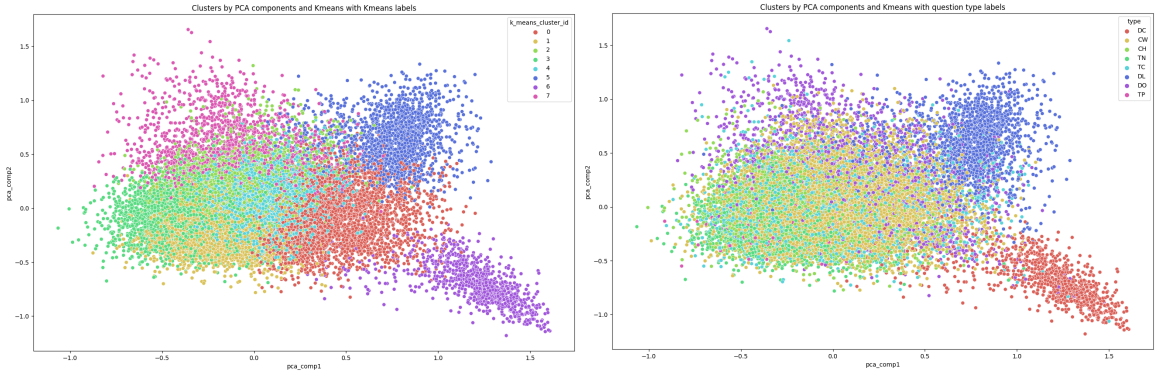
Xiao, J., Shang, X., Yao, A., and Chua, T.-S. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

A. Appendix

A.1. Embeddings plots for text embeddings fine-tuned through multimodal pre-training

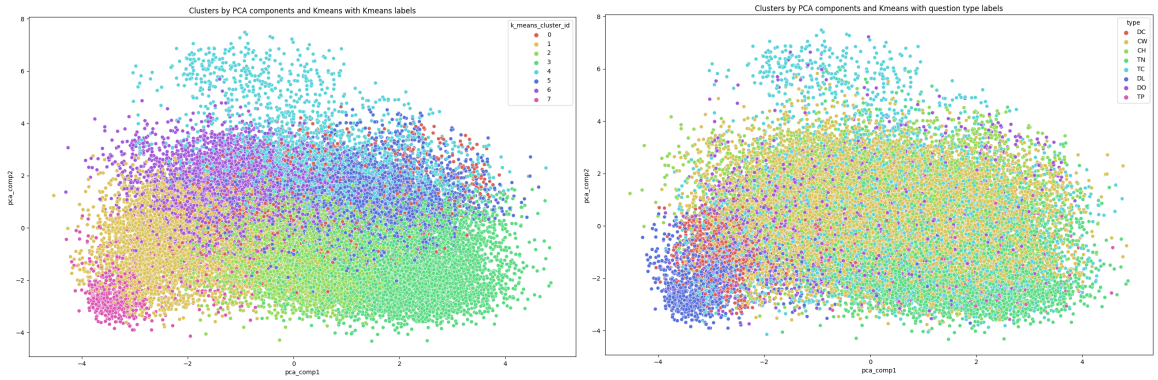
Here we present some more graphs for the multimodal fine-tuned text embedding representations.



(a) K-means assigned cluster ID

(b) Question type labels for K-means clusters

Figure 6. K-means clustering with $< q, a >$ GloVe Embeddings

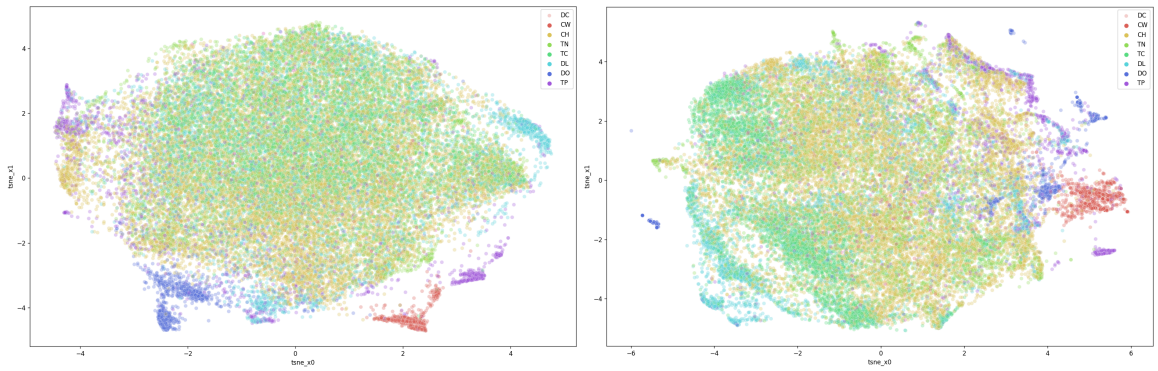


(a) K-means assigned cluster ID

(b) Question type labels for K-means clusters

Figure 7. K-means clustering with $< q, a >$ BERT Embeddings

A.2. t-SNE plots for text modality $< q, a >$ embeddings

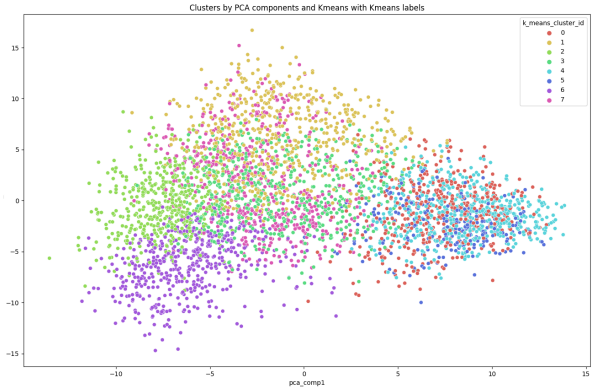


(a) $< q, a >$ GloVe Embeddings

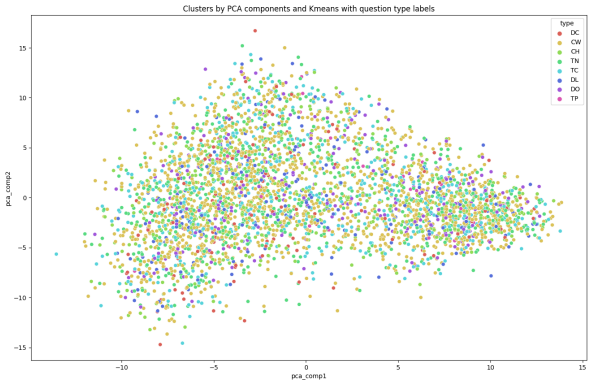
(b) $< q, a >$ BERT Embeddings

Figure 8. t-SNE plots for text modality $< q, a >$ embeddings

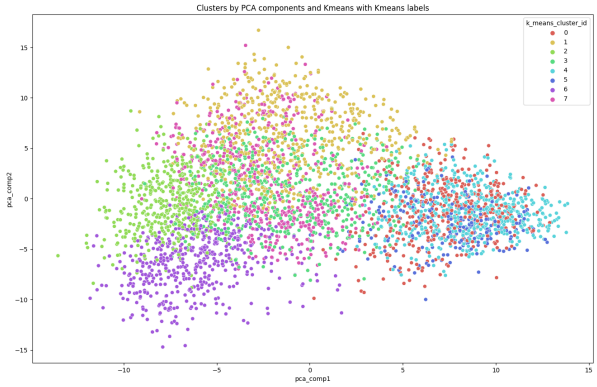
A.3. K-means clustering based on identified clusters and question types for the visual features of the videos



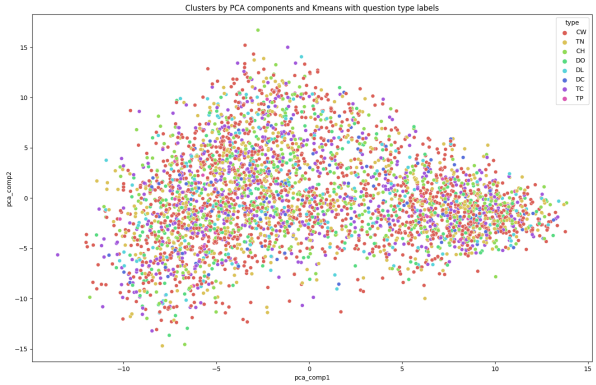
(a) K-means assigned cluster ID for first 3870 QA pairs



(b) Question type labels for K-means clusters first 3870 QA pairs



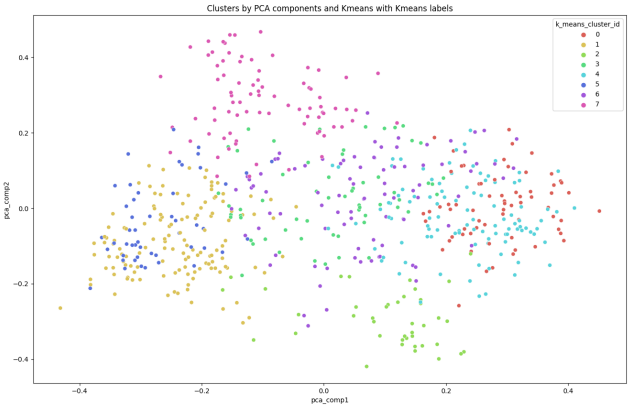
(c) K-means assigned cluster ID for next 3870 QA pairs



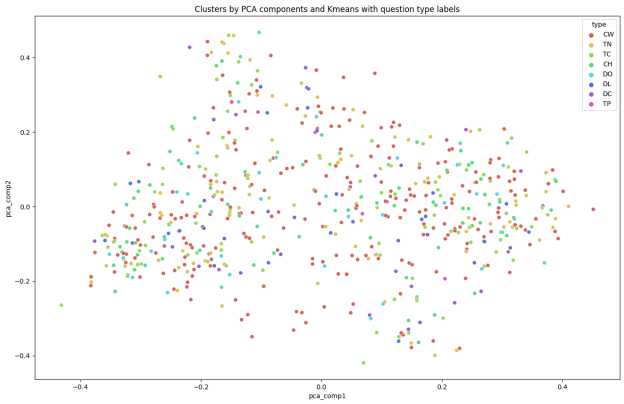
(d) Question-type labels for K-means clusters of next 3870 QA pairs

Figure 9. K-means clustering of Resnet and ResNeXt features from HCRN

A.4. S3D and ViT feature based clustering of the videos

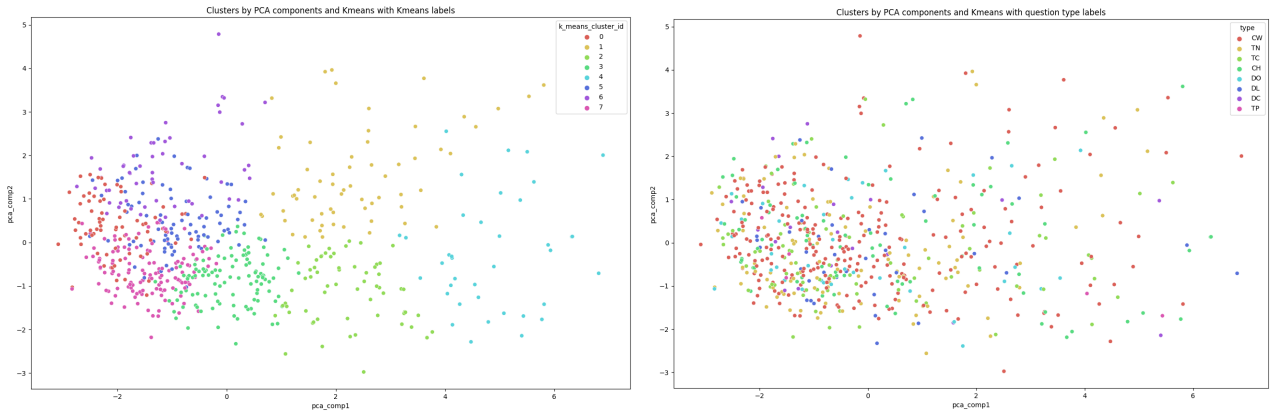


(a) K-means assigned cluster ID



(b) Question-type labels for K-means clusters

Figure 10. K-means clustering with S3D video features

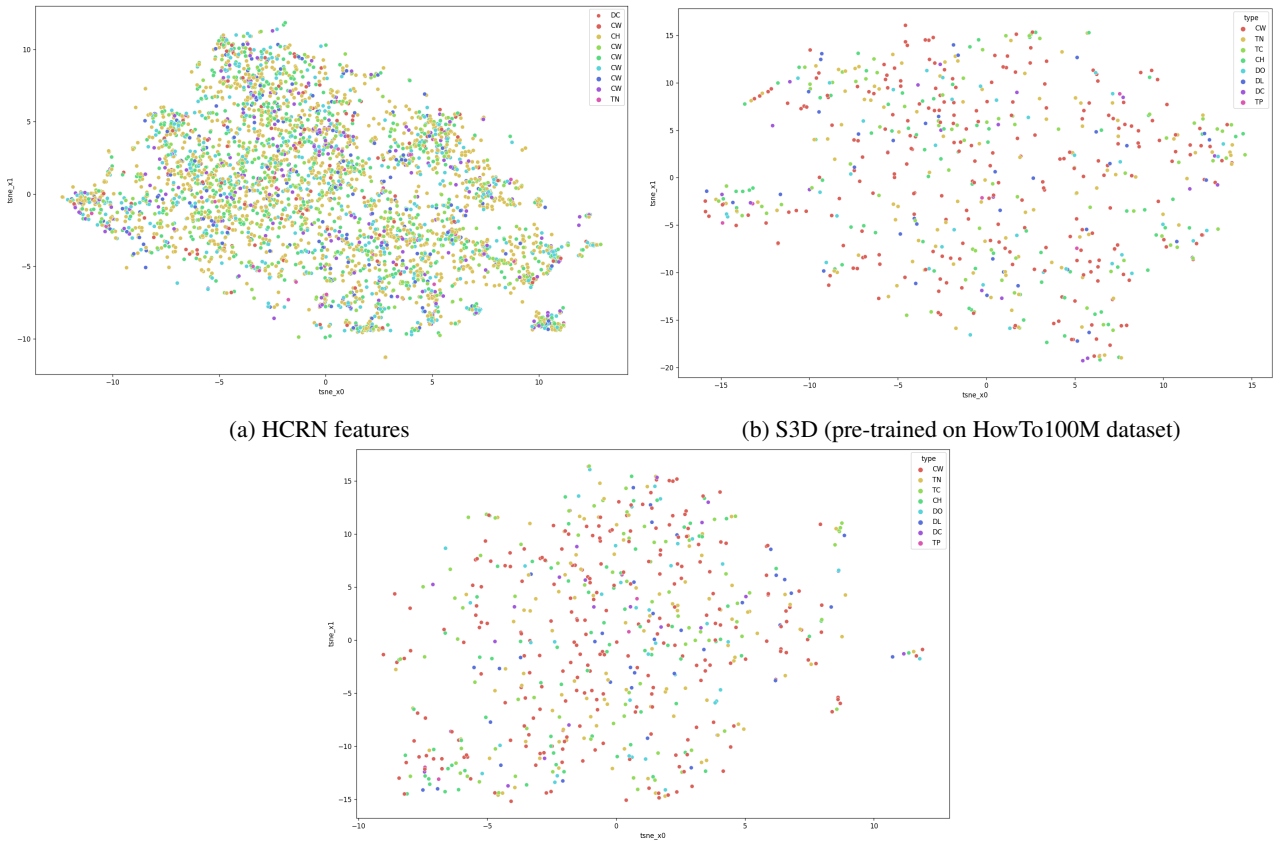


(a) K-means assigned cluster ID

(b) Question-type labels for K-means clusters

Figure 11. K-means clustering with VideoMAE (ViT-based) video features

A.5. t-SNE analysis of HCRN, S3D and VideoMAE baseline features for the videos



(a) HCRN features

(b) S3D (pre-trained on HowTo100M dataset)

(c) VideoMAE (ViT-based)

Figure 12. t-SNE on HCRN, S3D, and VideoMAE video features