Reliable Ambient Intelligence Through Large Language Models

Stanford CS224N Custom Project

Wei Dai
Department of Computer Science
Stanford University
dvd.ai@stanford.edu

Abstract

While computer vision has gained traction in medical applications, models specifically engineered for Intensive Care Unit (ICU) activities are limited and often lack clinical relevance. To tackle this challenge, we present Clinical Behavioral Atlas (CBA), a computer vision system that can identify 40 clinically relevant activity categories, 55 object categories, and three personnel categories solely through RGB video data. In addition, we develop a novel LLM-based model grounded on patient activity data for accurate video question answering and captioning. The specific activity grounding, fine-tuned on the CBA dataset, reduces hallucinations and allows the LLM to focus more on clinically relevant information, increasing its utility in the clinical domain.

External mentor: Ehsan Adeli, Alan Luo, Dev Dash, Mentor: Tony Wang

1 Introduction

The rapidly evolving field of computer vision, a specialized area within machine learning dedicated to algorithmic interpretation of visual data, has shown remarkable progress in various industries (Kamilaris and Prenafeta-Boldú, 2018; Silver et al., 2016; Yurtsever et al., 2020). Despite these advancements, its application in healthcare—particularly beyond the specialized fields of radiology and pathology (Shen et al., 2017)—remains largely unexplored. Implementing computer vision technologies in settings such as Intensive Care Units (ICUs) presents a compelling value proposition (Sathyanarayana et al., 2018; Tscholl et al., 2020). These sophisticated algorithms can offer a continuous, automated layer of visual data that fills what our team refers to as the "dark spaces of healthcare" (Haque et al., 2020). Operating tirelessly and without the limitations of human fatigue, these models not only facilitate real-time clinical decision-making but also integrate seamlessly into Electronic Medical Records (EMRs) for enhanced documentation. This enriched data repository can serve as the basis for both descriptive and quantitative analyses, revolutionizing how healthcare providers understand and engage with patient behaviors (Lloyd-Jukes et al., 2021).

Among the numerous aspects of video comprehension, Video Question Answering (VideoQA) has garnered a significant amount of attention, since it requires models to answer questions regarding a specific video segment, which necessitates a thorough grasp of the scene, relationships, and temporal changes depicted in the video Zhong et al. (2022); Jang et al. (2017); Xiao et al. (2021). However, current video models struggle to obtain a fine-grained understanding of the video scene Wang et al. (2023). This issue is particularly detrimental in the video domain, where a miss in the scene understanding may cause life injuries. As a result, developing a clinical-oriented model for video understanding requires specialized attention to clinically relevant details.

Building on our prior research in ambient intelligence, our study aims to address these challenges by developing a specialized video question-answering model for ICUs. Our model employs an activity recognizer, which is designed to automatically recognize a broad spectrum of clinically meaningful activities in the ICU, encompassing activities of daily living (ADLs), preventive measures, diagnostic

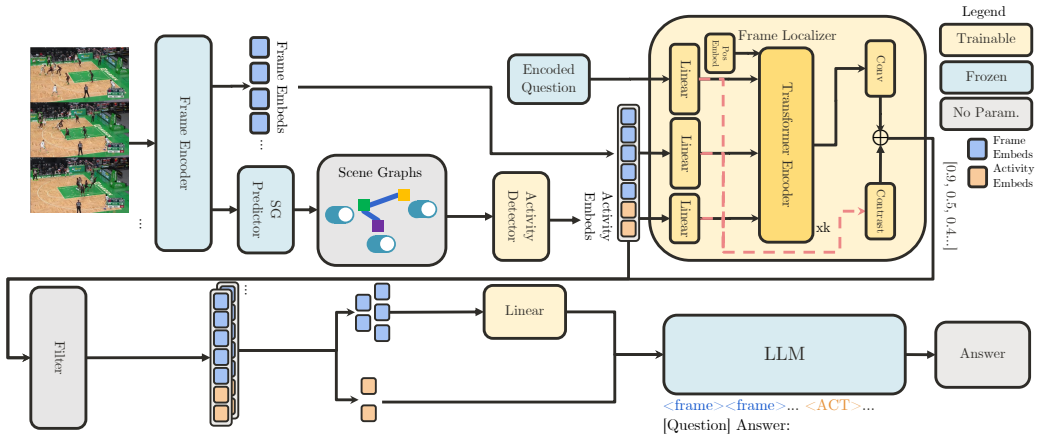

Figure 1: **Model Architecture.**

procedures, therapeutic procedures, and other clinically important bedside physical activities. The resulting model not only achieves strong performance but also reveals previously inaccessible insights into care delivery, which aids clinical workers in achieving better patient outcomes. The backbone of our model is a frozen LLM, which provides better reasoning ability and is able to extract the complex semantic information provided by the activity recognizer. We hope our design can serve as a paradigm for clinical-oriented model training, aiding clinical workers to give better care in the future.

2 Related Work

2.1 Clinical Activity and Behavior Recognition

Traditional non-vision-based approaches to identifying clinically relevant behaviors often fall short due to their rudimentary nature or narrowly defined use cases (Srigley et al., 2014). For instance, current systems, such as bed motion alarms, are limited in scope and do not provide a comprehensive understanding of patient activities. Moreover, these systems are plagued by workforce limitations (Patel et al., 2018), affecting the precision and timeliness of patient data capture.

2.2 Video Question Answering

The effectiveness of machine learning models is profoundly impacted by the data on which they are trained. Key datasets such as MovieQA (Tapaswi et al., 2016), MSRVTT-QA, and MSVD-QA (Xu et al., 2017) have played a crucial role in propelling the field of video question answering forward (Peng et al., 2022; Lei et al., 2021; Yang et al., 2021; Jiang and Han, 2020). These datasets, however, primarily feature brief video segments accompanied by straightforward questions, which restricts the progression of models toward achieving a comprehensive understanding of videos. The TGIF-QA dataset (Jang et al., 2017) marked a pivotal shift by evaluating models on spatial-temporal reasoning across a vast collection of animated GIFs, thereby enhancing the temporal reasoning capabilities of models (Gao et al., 2018; Fan et al., 2019). Nonetheless, current models fail to capture fine-grained information that is pertinent to clinical use cases.

3 Approach

As illustrated in Figure 1, our model fuses multimodal video frame features with language for advanced video understanding. A detailed illustration of each component is described below.

Frame Encoder. To generate patch image features $X \in \mathbb{R}^{L_{patch} \times d_v}$, object bounding box predictions $B = b_1, \ldots, b_n$, object class predictions $O = o_1, \ldots, o_n$, we employ EVA-02 (Fang et al., 2023), a ViT-based image encoder with 304M parameters. In addition, we modify it by incorporating an additional attribute head to classify attributes for each entity. Leveraging the pre-trained weights from the original research, we fine-tune this model on the CBA dataset.

Scene Graph Predictor and Activity Recognizer. Our study introduces a hierarchical approach to the recognition of clinical activities. The bottom level builds scene graphs using modules for entity detection, entity tracking, keypoint detection, and attribute classification. Specifically, when an actor is identified by the backbone frame encoder, we employ ViTPose (Xu et al., 2022) to detect 133 distinct human body keypoints. We derive relationship information using a heuristic method that measures pairwise distances between certain bounding boxes or keypoints. Lastly, we have developed a Multi-Class Multi-Object Tracking (MCMOT) algorithm based on ByteTrack (Zhang et al., 2022) to correlate entity instances across different frames. The top level classifies the presence of activity classes at each time step in a video using the raw video data and the extracted scene graphs. We use the GINE (Xu et al., 2018) graph neural network to encode the scene graph at the target frame. We concatenate the two extracted encodings and feed them into a MLP layer to make the final prediction.

Frame Localizer. The frame localization method, utilizing the UniVTG structure as proposed by Lin et al. (Lin et al., 2023), adopts a dual strategy of alignment and contrast for processing frame and activity representations. In its training phase, it assigns binary labels f_i to frames, where $f_i = 1$ indicates a clip is part of the foreground, alongside a saliency score $s_i \in [-1, 1]$ that denotes its importance to the posed question. It converts the question into a series of query tokens $\mathbf{Q} \in \mathbb{R}^{n \times d_t}$. Each frame's embedding $\mathbf{X}$ and activity embedding $\mathbf{S}$ are processed through individual linear transformations:

$$\mathbf{x}_i = \frac{1}{|\mathbf{X_i}|} \sum_{j=1}^{|\mathbf{X_i}|} \mathbf{X_i} \mathbf{W}_{xs}, \quad \mathbf{s}_i = \frac{1}{|\mathbf{S_i}|} \sum_{j=1}^{|\mathbf{S_i}|} \mathbf{S_i} \mathbf{W}_{ss},$$

where $\mathbf{W}xf, \mathbf{W}_{sf}$ represent matrices that can be adjusted. The condensed frame and activity embeddings are then merged to create the video frame embedding $\mathbf{X}_v = \mathbf{x}_1, \ldots, \mathbf{x}_n$ and the video activity embedding $\mathbf{S}_v = \mathbf{s}_1, \ldots, \mathbf{s}n$ for videos of length n. For alignment, positional and type embeddings are added to each modality: $\mathbf{X}'_v = \mathbf{X}_v + \mathbf{E}_X^{pos} + \mathbf{E}_X^{type}; \mathbf{Q}'_v = \mathbf{Q}_v + \mathbf{E}_Q^{pos} + \mathbf{E}_Q^{type}; \mathbf{S}'_v = \mathbf{S}_v + \mathbf{E}_S^{pos} + \mathbf{E}_S^{type}$. Then, the embeddings concatenated into $\mathbf{Z}_0 = [\mathbf{X}'_v; \mathbf{S}'_v; \mathbf{Q}'_v]$. These embeddings are then combined into $\mathbf{Z}0 = [\mathbf{X}v'; \mathbf{S}v'; \mathbf{Q}v']$. This unified representation $\mathbf{Z}0$ is processed through a sequence of k transformer encoders, each consisting of a multi-head self-attention mechanism (MHSA) and a linear layer. For encoder layer i with m attention heads, we have

$$\mathbf{k}_{i,m} = \text{softmax}\left(\frac{\mathbf{W}_Q^{i,m}\mathbf{Z}_{i-1}(\mathbf{W}_K^{i,m}\mathbf{Z}_{i-1})^T}{\sqrt{d_k^i}}\right), \mathbf{h}_{i,m} = \mathbf{k}_{i,m}\mathbf{W}_V^{i,m}\mathbf{Z}_{i-1},$$
$$\mathbf{Z}_i = (||_{m=1}^{M}\mathbf{h}_{i,m})\mathbf{W}_O^i,$$

where $\mathbf{W}_O^i, \mathbf{W}_Q^{i,m}, \mathbf{W}_K^{i,m}, \mathbf{W}_V^{i,m}$ are learnable parameters. In the end, we take out the question tokens to get $\mathbf{Z}'_k$. The score for the alignment route is then

$$\hat{\mathbf{f}} = \sigma(\text{Conv}(\mathbf{Z}'_k)), \tag{1}$$

where σ is a sigmoid activation, and Conv is a set of convolutional layers that outputs $\hat{\mathbf{f}} \in \mathbb{R}^n = \{\hat{f}_1, \ldots, \hat{f}_n\}$. This score signals whether the given frame corresponds to a foreground clip. This route is supervised by the cross entropy loss between the predicted label $\hat{\mathbf{f}}_a$ and the ground truth label f_a:

$$\mathcal{L}_a = \sum_{i=1}^{n} -\left(f_i \log \hat{f}_i + (1 - f_i)\log(1 - \hat{f}_i)\right), \tag{2}$$

where s_i is the ground truth relevance at frame i.

The saliency score is learned through contrastive learning. In this route, a linear layer is first used to project each embedding into the same embedding space. Then, the saliency score $\hat{s}_c$ is obtained through the sum of the pair-wise similarity score between the activity embedding $\mathbf{S}_v = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, frame embedding $\mathbf{X}_v = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, and question embedding $\mathbf{Q}$:

$$\hat{s}_{c,i} = \frac{\mathbf{x}_i^T\mathbf{Q}}{||\mathbf{x}_i||_2||\mathbf{Q}||_2} + \frac{\mathbf{s}_i^T\mathbf{Q}}{||\mathbf{s}_i||_2||\mathbf{Q}||_2}. \tag{3}$$

This score is supervised through two losses: intra-video and inter-video contrastive learning loss. For intra-video contrastive learning loss, we randomly sample a positive clip at index p with $f_p = 1$ and

$s_p > 0$, and negative samples $N = \{j | 1 \leq j < p, s_j < s_p\}$. Given the saliency prediction $\hat{s}_j, \hat{s}_j$, the intra-video loss is calculated as

$$\mathcal{L}_s^{\text{intra}} = -\log \frac{\exp(\hat{s}_p/\tau)}{\exp(\hat{s}_p/\tau) + \sum_{j \in N} \exp(\hat{s}_j/\tau)}, \tag{4}$$

where τ is the temperature chosen manually. The inter-video loss takes other videos $k \in N'$ within the batch as negative samples

$$\mathcal{L}_s^{\text{inter}} = -\log \frac{\exp(\hat{s}_p/\tau)}{\sum_{k \in B} \exp(\hat{s}_p^k/\tau)}. \tag{5}$$

We take a weighted sum to obtain the final loss:

$$\mathcal{L} = \lambda_a \mathcal{L}_a + \lambda_{intra} \mathcal{L}_s^{\text{intra}} + \lambda_{inter} \mathcal{L}_s^{\text{inter}}, \tag{6}$$

where $\lambda_a, \lambda_{inter}, \lambda_{intra}$ are hyperparameters setting the weight for each loss. Similarly, another weighted sum is performed to obtain the final score $\hat{r}_i$:

$$\hat{r}_i = w_f \hat{f}_i + w_s \hat{s}_i. \tag{7}$$

For $\hat{r}_i$, though, the weights w_f, w_s are learned during the training process. The frames are ranked based on $\hat{r}_i$, and only the top k frames are input into the Q-Formers in the next stage. For datasets with ground truth interval annotation, the localizer is directly tuned on the ground truth labels. For datasets without ground truth labels, pseudo labels f_i', s_i' are generated to fine-tune the localizer. Specifically, for each frame i with answer prediction $y, \hat{y}$ and frame selection threshold r_θ,

$$f_i', s_i' = \begin{cases} 1, 1 & \text{if } (y = \hat{y} \wedge \hat{r}_i > r_\theta) \\ & \quad \vee (y \neq \hat{y} \wedge \hat{r}_i < r_\theta) \\ 0, -1 & \text{Otherwise} \end{cases}. \tag{8}$$

In other words, we encourage the localizer to make the same prediction if such prediction gives the correct answer while encouraging the model to make a different prediction when the selected frames fail to provide the correct answer.

LLM. We implement the LLM backbone based on the LLaVA-1.5 architecture with Mistral-7b backbone. A linear projection is then used to project the embedding into the LLM embedding space. Finally, the LLM takes both activity embeddings and the frame embeddings as inputs and an LLM inference is performed to obtain the final answer.

4 Experiments

4.1 Data

We evaluate the model on our in-house clinical dataset called Clinical Behavioral Atlas (CBA). A detailed description of the dataset is included below.

Data Collection. Our dataset comprises a continuous collection of video data obtained from two Intensive Care Units (ICUs) at Stanford Hospital, covering the period from November 2021 to November 2022. For data acquisition, a network of 16 sensors was employed across eight ICU rooms: one room in the Neuro-ICU and the remaining seven in the Medical-ICU. Two cameras were deployed in each room. Our sensors record the data with a resolution of 1920x1080 at 15 frames per second.

Taxonomy. Overall, we have chosen 40 activity classes categorized under five superclasses: preventive measures, activities of daily living, diagnostic procedures, therapeutic procedures, and personnel. At a more detailed level, we have included 55 object categories in our taxonomy.

The activity classes are motivated by five clinical bundles: the ABCDEF Bundle, the Deep Vein Thrombosis (DVT) Prophylaxis Bundle, the Hospital-Acquired Pressure Injury (HAPI) Prevention Bundle, the Healthcare-Associated Infections (HAI) Prevention Bundle, and the Ventilator-Associated Pneumonia (VAP) Prevention Bundle. The ABCDEF bundle is a comprehensive approach to patient care focusing on Assessing and managing pain, Both spontaneous awakening and breathing trials,

Model	Describe		Temporal		Binary	Aggregated	
	Acc	BLEU	Acc	BLEU	Acc	Acc	BLEU
BLIP-2 (Li et al., 2023)	10.4	0.043	39.9	0.066	69.9	44.0	0.048
SeViLA Yu et al. (2023)	21.1	0.046	36.6	0.064	71.7	47.4	0.049
LLaVA-1.5 (Liu et al., 2024)	**47.5**	0.103	44.3	0.158	58.8	51.8	0.127
Ours	42.2	**0.113**	**59.6**	**0.208**	**88.0**	**67.0**	**0.154**

Table 1: **Performance of our model against various baselines on CBA-QA.** The best metrics in each column are bolded.

Choice of sedation, Delirium monitoring and management, Early mobility and exercise, and Family engagement and empowerment, to improve outcomes for critically ill patients. Additionally, aligned with nursing documentation practices, we have incorporated categories from a nursing Epic Flowsheet. Lastly, our taxonomy encompasses categories of clinical equipment used in patient care.

Data Annotation. Data annotation was carried out by our research team in collaboration with contracted nurses, patient safety monitors, and an annotation vendor. In the first phase, researchers, clinicians, and other healthcare professionals selected clinically relevant video clips and annotated activities at the clip level. In the second phase, non-clinical annotators from the annotation vendor carried out detailed annotations. They annotated bounding boxes around entities, labeled their predicates, and tracked these entities throughout the clip. All annotations used data labels from our study taxonomy.

Dataset Splits. As we used video footage from 8 rooms, we split the data so that the testing and validation dataset each contains footage from one room, and the training dataset contains footage from the remaining 6 rooms. This way, no patient will appear in both training and testing dataset, ensuring maximum separation between them.

CBA-QA Dataset. The CBA-QA dataset is generated from the CBA dataset using the original footage and annotations. It consists of 52,426 question-answer pairs. Out of all questions, 16,924 of them ask the model to describe a video, 12,576 demand temporal understanding, and 22,926 involve binary questions on patient states.

4.2 Baselines

We compare our method against multiple LLM-based video baselines, including the vanilla LLaVA (Liu et al., 2024), BLIP-2 (Li et al., 2023) and SeViLA (Yu et al., 2023). While their backbone LLMs are of different dimensions, we keep their size similar in the 7b-13b range.

4.3 Evaluation method

As the task is video question answering, the model outputs answers given a video and a natural language question. The model is evaluated on two metrics: accuracy and BLEU score. Specifically, with test dataset $\mathbf{Q}$, the accuracy of the prediction $\hat{q}$ with respect to ground truth q is given by: $acc = \frac{1}{|\mathbf{Q}|}\sum_{\mathbf{Q}}\frac{1}{|q|}\sum_{i=1}^{min(|\hat{q}|,|q|)}\mathbf{I}[\hat{q}_i = q_i]$. For BLEU score, we used the vanilla computation method (Papineni et al., 2002) with ORANGE smoothing (Lin and Och, 2004).

4.4 Experimental details

The backbone LLM, based on the Mistral-7b structure, is kept frozen. The frame encoder is trained on the bounding box and attributes annotation data of the CBA dataset on 4 NVIDIA A100 GPUs. Within the frame encoder, the bounding box predictor has a hidden dimension of 1024, and the backbone feature extractor is a standard vision transformer (ViT) with an embedding dimension of 768. The model is trained with a learning rate of 5×10^{-7}, a batch size of 40 for a total of 10 epochs.

4.5 Results

The experimental result is included in Table 1. As shown in the table, our model performs the best in most of the categories, with an increase in accuracy of 15.7% and an increase in BLEU of 21.3%.

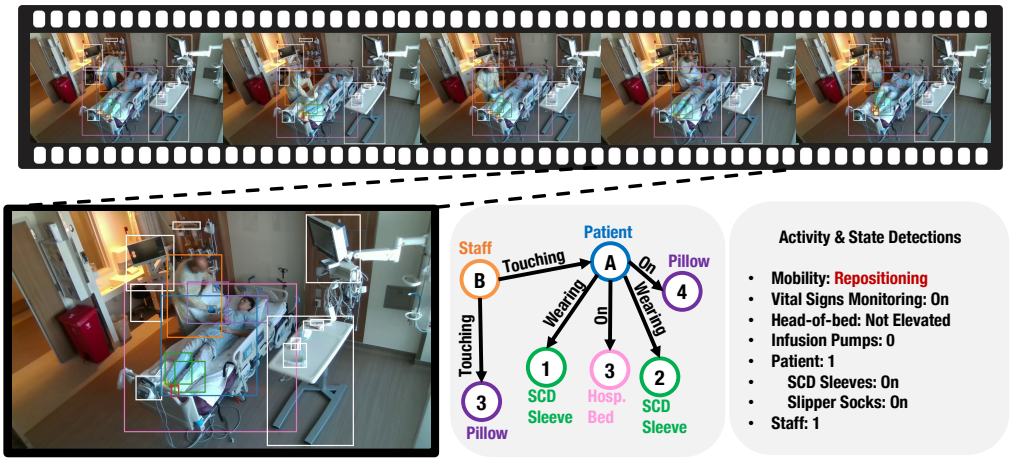

Figure 2: Visualization of the activity recognizer on simulation data. The simulation data does not contain patient information, and all people in the scene are researchers in the project.

This result signals the efficacy of the activity recognizer in helping the model achieve a better scene understanding in the clinical domain. The most prominent performance boost is in the binary category, with 29.2% better accuracy. This demonstrates how an accurate temporal grounding can help a model to understand a scene better.

The model under examination did not achieve performance parity with LLaVA regarding the accuracy of descriptive questions despite exhibiting superior BLEU scores. Upon thorough analysis, it is posited that this discrepancy arises from the challenges inherent in integrating additional activity grounding, which necessitates extra contextual information. This requirement appears to encumber the reasoning capabilities and the capacity to generate cohesive descriptive paragraphs in smaller models. We hypothesize that transitioning to larger language models, specifically from the 7b to more expansive language architectures, may ameliorate this issue.

Our observations further reveal that all models under consideration exhibit relatively low BLEU scores. This phenomenon can primarily be attributed to the inherent variation in wording styles among LLMs, which may not align with the style of the reference texts. Consequently, despite the accurate and comprehensive representation of factual content, the models are assigned low BLEU scores. This discrepancy suggests that the BLEU metric may not serve as an effective tool for evaluating language model performance in the context of LLMs. Nonetheless, alternative evaluation metrics, such as the GPT rating, have been subject to criticism for their lack of consistency across different model generations.

5 Analysis

To investigate the performance of the activity recognizer module, we visualized both the final and intermediate outputs of the activity recognizer in Figure 2. The source video in the figure is collected through simulation, and all people in the scene are researchers in the project. On the object level, the model can identify large objects as well as tiny objects like the paper cup. It, however, failed to identify one occluded slipper sock and could only identify one. On the scene graph level, the model can identify a variety of relationships, including *touching*, *wearing* and *on*. The relationships, however, are limited to spatial relationships. Complex semantic relationships, such as *talking to*, are areas the model comes short of. On the activity level, the model accurately extracts the information from the scene graph, identifying that the patient in the scene is being repositioned. In addition, the model identifies that there is one patient and one staff member on the scene, with the patient having SCD sleeves and slipper socks on the legs and feet. This shows the robustness of our model: Even though some objects, like one slipper sock, are not identified, the model is still able to obtain an accurate activity and state prediction through partial data.

6 Conclusion

In this work, we introduce a video question-answering model for clinical understanding. Our model employs an activity recognizer that accurately extracts clinical-related information from a video. The model sets a new SoTA in the CBA-QA dataset, a video dataset designed for clinical question answering. We hope the work can serve as a paradigm to guide further clinical model development in the area, helping large language models to reach clinical level performance in the future.

The predictive accuracy of the model, contingent upon activity information, is noteworthy; however, the construction of the activity recognizer remains a manual process. Furthermore, the selection of pertinent information necessitates pre-existing domain-specific knowledge. For future endeavors, it is posited that a more adaptable approach, allowing the model itself to determine the relevant information for inclusion in the grounding data, could facilitate the development of a more universally applicable solution that does not require prior clinical expertise.

References

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, pages 1999–2007. Computer Vision Foundation / IEEE.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*.

Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *CVPR*, pages 6576–6585. Computer Vision Foundation / IEEE Computer Society.

Albert Haque, Arnold Milstein, and Li Fei-Fei. 2020. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature*, 585(7824):193–202.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. 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.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, pages 11109–11116. AAAI Press.

Andreas Kamilaris and Francesc X Prenafeta-Boldú. 2018. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341. Computer Vision Foundation / IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. Univtg: Towards unified video-language temporal grounding. In *CVPR*, pages 2794–2804.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Hugh Lloyd-Jukes, Oliver John Gibson, Tracey Wrench, Ade Odunlade, and Lionel Tarassenko. 2021. Vision-based patient monitoring and management in mental health settings. *Journal of Clinical Engineering*, 46(1):36–43.

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

Rikinkumar S Patel, Ramya Bachu, Archana Adikey, Meryem Malik, and Mansi Shah. 2018. Factors related to physician burnout and its consequences: a review. *Behavioral sciences*, 8(11):98.

Min Peng, Chongyang Wang, Yuan Gao, Yu Shi, and Xiang-Dong Zhou. 2022. Multilevel hierarchical network with multiscale sampling for video question answering. In *IJCAI*, pages 1276–1282. ijcai.org.

Supriya Sathyanarayana, Ravi Kumar Satzoda, Suchitra Sathyanarayana, and Srikanthan Thambipillai. 2018. Vision-based patient monitoring: a comprehensive review of algorithms and technologies. *Journal of Ambient Intelligence and Humanized Computing*, 9:225–251.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. 2017. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Jocelyn A Srigley, Colin D Furness, G Ross Baker, and Michael Gardam. 2014. Quantification of the hawthorne effect in hand hygiene compliance monitoring using an electronic monitoring system: a retrospective cohort study. *BMJ quality & safety*, 23(12):974–980.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640. IEEE Computer Society.

David Werner Tscholl, Julian Rössler, Sadiq Said, Alexander Kaserer, Donat Rudolf Spahn, and Christoph Beat Nöthiger. 2020. Situation awareness-oriented patient monitoring with visual patient technology: a qualitative review of the primary research. *Sensors*, 20(7):2112.

Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. 2023. VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5036–5048, Toronto, Canada. Association for Computational Linguistics.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1666–1677. IEEE.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6439–6455, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.