Natural Language Queries in Egocentric Videos

Yasaman Noshirvanbaboli
S327167@studenti.polito.it

Seyed Mohammad Sheikh Ahmadi
mohammad.sheikh@studenti.polito.it

Amirhossein Lotf Ranaei
amirhossein.lotfranaei@studenti.polito.it

Abstract

Egocentric vision, which captures videos from the first-person perspective using head-mounted cameras, provides a unique viewpoint of human interactions and daily activities. The release of large-scale egocentric datasets, such as EPIC-KITCHENS and Ego4D, has spurred significant research in this field. Traditional tasks include action recognition and anticipation, but natural language queries (NLQs) now allow users to interact with video data using conversational language. This project explores the challenges and methodologies of NLQs in egocentric videos, focusing on the Ego4D NLQ benchmark, which involves temporal segment prediction to identify the exact moment a query is answered in a video. We evaluated two models, VSLNet and VSLBase, using features extracted from Omnivore and EgoVLP, two advanced feature extraction models. Our experiments were designed to compare the effectiveness of these features and models in addressing the NLQ task. Extensive hyperparameter tuning was performed to ensure robust comparisons. The results demonstrated that the combination of EgoVLP features and the VSLNet model with a shared encoder configuration achieved the highest performance, with an overall mIoU of 7.63. Additionally, we implemented an extension to the project by developing a natural language query generation system using Google's open-source model, Gemma 2B EN (ar5iv). This system was fine-tuned to generate NLQs from descriptive sentences, enhancing the interaction with egocentric video data. Our findings highlight the importance of specialized features and sophisticated models for NLQs in egocentric vision, providing insights for future research in this domain. The results have significant implications for personal assistance, accessibility, behavioral analysis, and education. For the complete code related to this project, please visit the public GitHub repository: https://github.com/smsag99/episodic-memory.

1. Introduction

Egocentric vision, which involves capturing videos from the first-person perspective using cameras mounted on the user's head, provides a unique and privileged viewpoint of human interactions and daily activities [1]. The release of large-scale egocentric datasets, such as EPIC-KITCHENS and Ego4D, has spurred significant research in this field [2]. Traditional tasks include action recognition and anticipation, but natural language queries (NLQs) now allow users to interact with video data using conversational language [3].

The Ego4D NLQ benchmark is a pioneering effort, defining the task as temporal segment prediction where models identify the exact moment a query is answered in a video [6]. This involves integrating video and text modalities, managing long temporal ranges, and ensuring precise temporal alignment [7].

This report explores the challenges and methodologies of NLQs in egocentric videos, discussing technological frameworks like LifelongMemory that integrate large language models (LLMs) with pre-trained NLQ models to enhance performance. It also examines the broader implications of this technology in personal assistance, accessibility, behavioral analysis, and education, aiming to provide a comprehensive understanding of the current state and future directions of NLQs in egocentric vision.

2. Related Work

The domain of egocentric vision has seen significant advancements, driven by the proliferation of large-scale datasets and sophisticated models capable of understanding and processing first-person video data. This section reviews key contributions to the field, focusing on seminal works that have laid the foundation for current research in natural language queries (NLQs) within egocentric videos.

Grauman *et al.*, 2022 [1] introduced the Ego4D dataset, a monumental effort capturing over 3,000 hours of egocentric video from various geographical and cultural contexts. The dataset's scale and diversity make it a critical resource for

developing models capable of understanding complex, real-world activities from a first-person perspective. Ego4D also sets the stage for numerous benchmarks, including action recognition, anticipation, and natural language queries, fostering advancements in the integration of computer vision and NLP (ar5iv).

Gao *et al.*, 2017 [3] introduced the TALL framework, which focuses on temporal activity localization in videos through language queries. This work pioneers the approach of using natural language to specify temporal segments within videos, a concept crucial for developing NLQ systems in egocentric vision. TALL combines visual and textual features to accurately identify and localize activities based on user queries, demonstrating the feasibility and effectiveness of such models (ar5iv).

Most similar to our work is Zhang *et al.*, 2020 [4], who proposed a span-based network for localizing moments in videos using natural language descriptions. This model significantly improves the precision of temporal segment prediction by treating the localization task as a span prediction problem, akin to question answering in NLP. The approach enhances the ability to handle complex queries and accurately align them with corresponding video segments.

Zhang *et al.*'s work on VSLNet and VSLBase demonstrates the effectiveness of span-based methods for NLQs in videos, particularly in handling the nuanced and context-dependent nature of egocentric video content. The ability of these models to accurately localize relevant video segments based on natural language queries makes them highly relevant to our project [4], which aims to enhance the interaction and usability of egocentric vision systems [10].

Feichtenhofer *et al.*, 2019 [5] introduced SlowFast networks, a dual-pathway model designed to capture both slow and fast visual dynamics in videos. This architecture is particularly well-suited for egocentric video analysis, where activities can range from subtle, slow movements to rapid actions. The SlowFast networks' ability to process varying temporal resolutions enhances the robustness and accuracy of action recognition and NLQ systems (ar5iv).

3. Methodology

3.1. Ego4D Dataset

The dataset employed in this project is Ego4D, a comprehensive egocentric video dataset capturing a wide range of daily life activities from a first-person perspective [4]. Ego4D is distinguished by its scale and diversity, encompassing over 3,000 hours of video footage collected from multiple geographical locations and cultural contexts [1]. The dataset provides a rich resource for developing and benchmarking egocentric vision models.

This project focuses on the Natural Language Queries (NLQ) benchmark within Ego4D. The NLQ task involves predicting the temporal segments of videos where answers to natural language queries can be found. This task poses a significant challenge due to the need for a fine-grained understanding of both video content and query semantics [6].

3.2. Annotations and Narrations

The Ego4D dataset provides extensive annotations and narrations that are crucial for various tasks, including the NLQ benchmark. These narrations describe the actions and interactions occurring within the video, offering a rich source of contextual information. Annotations include temporal boundaries for activities, object interactions, and associated metadata. This detailed labelling facilitates the training of models to understand and respond to natural language queries accurately.

The annotations and narrations from Ego4D play a pivotal role in training and evaluating the models. They provide the necessary ground truth for understanding the context of queries and enhancing the model's ability to generate accurate responses.

3.3. Feature Extraction

For this project, we utilized two distinct sets of features extracted from the Ego4D videos:

1. **Omnivore Features:** Omnivore is a model designed to handle multiple visual modalities, including images, videos, and depth data. It leverages a unified architecture to extract rich and versatile visual features [11]. For the NLQ task, we used video features derived from the Omnivore model, which are particularly adept at capturing detailed visual information from egocentric perspectives.

2. **EgoVLP Features:** EgoVLP (Egocentric Video-Language Pre-training) is a contrastive model pre-trained on a subset of Ego4D videos using weak supervision from free-form textual narrations. This model aligns video segments with corresponding textual descriptions, making it highly effective for tasks that involve video-language interaction. The features extracted from EgoVLP are designed to bridge the gap between visual content and language queries, enhancing the model's ability to locate relevant video segments [6].

3.4. Models

We evaluated two different models for the NLQ task, each with its unique architecture and capabilities:

1. **VSLNet:** VSLNet (Video Span Localization Network) builds upon the concept of span prediction widely used in natural language processing [1, 4]. VSLNet treats

the task of video moment localization as a question-answering problem [8], where the start and end times of the relevant video segment are predicted in response to a natural language query [4].

2. **VSLBase:** VSLBase is a simplified version of VSLNet [4]. It uses a less complex architecture but follows a similar principle of joint video-query encoding and span prediction [11]. Despite its simplicity, VSLBase is trained on the same datasets and evaluated using the same metrics as VSLNet, providing valuable insights into the performance trade-offs between model complexity and localization accuracy [4, 10, 11].

3.5. Experimental Setup

To comprehensively evaluate the performance of the models, we designed experiments under four different conditions, combining the two models with the two feature sets:

1. **Omnivore + VSLBase:** Investigated the impact of using Omnivore features with a simpler model architecture.

2. **Omnivore + VSLNet:** Expected to improve temporal localization accuracy due to enhanced attention mechanisms.

3. **EgoVLP + VSLBase:** Investigated the impact of using features specifically pre-trained on egocentric video-language tasks.

4. **EgoVLP + VSLNet:** Hypothesized to achieve the highest accuracy by combining advanced features with an improved model architecture.

Each combination represents a unique configuration of features and model architecture, enabling us to analyze the interplay between different types of visual information and model capabilities.

3.6. Hyperparameter Tuning

We conducted an extensive hyperparameter search for the initial condition (VSLNet with Omnivore Features) to ensure robust and fair comparisons across different conditions. We tested seven different configurations and identified the optimal hyperparameters as follows:

- **Batch Size:** 16
- **Hidden Dimension (DIM):** 128
- **Number of Epochs (NUM_EPOCH):** 50
- **Maximum Position Length (MAX_POS_LEN):** 16
- **Initial Learning Rate (INIT_LR):** 0.0025

These hyperparameters were then consistently applied to the other three conditions, ensuring that variations in performance could be attributed to the features and model architecture rather than differences in training settings. The training and evaluation processes for the models involved several key steps to ensure rigorous assessment and reliable results.

3.7. Training Process

Feature Extraction: Utilizing the pre-extracted Omnivore and EgoVLP features, we ensured a standardized feature extraction process across all conditions. This step involved applying the Omnivore and EgoVLP models to the video data to obtain the respective feature representations [6, 11].

3.8. Evaluation Process

The performance of the models was evaluated using the following metrics:

1. **Top-k Recall:** This metric measures the proportion of correctly identified temporal segments among the top-k predictions. It provides insight into the model's ability to retrieve relevant segments from the video [12].

2. **Temporal Intersection over Union (tIoU):** tIoU measures the overlap between the predicted and ground truth temporal segments. It provides a robust evaluation of the accuracy of the model's temporal localization capabilities [3].

We compared the results of our models with the official baseline results provided in the Ego4D paper, which were trained on SlowFast features. This comparison allowed us to benchmark our models against established performance metrics and assess the effectiveness of our feature sets and model architectures.

3.9. Results Analysis

The primary objective of this project was to evaluate the effectiveness of different features (Omnivore vs. EgoVLP) and model architectures (VSLBase vs. VSLNet) in addressing the NLQ task. The analysis focused on:

1. **Impact of the Query Guided Highlighter Module:** By comparing the performance of VSLNet and VSLBase, we aimed to assess the contribution of the Query Guided Highlighter module in enhancing the model's precision and recall in temporal segment localization.

2. **Effectiveness of Omnivore and EgoVLP Features:** The comparison between Omnivore and EgoVLP features provided insights into the relative strengths of

each feature set in capturing relevant visual information for the NLQ task. This analysis helped identify which feature set offered better alignment with natural language queries.

3. **Overall Model Performance:** We analyzed the overall accuracy and robustness of each model-feature combination in predicting temporal segments. This involved a detailed examination of top-k recall and tIoU metrics across different configurations, providing a comprehensive evaluation of model performance.

4. Experiment

4.1. Experimental Procedure

The following steps were undertaken to evaluate the performance of the VSLBase and VSLNet models using two different feature sets (Omnivore and EgoVLP) on the Ego4D NLQ benchmark:

1. **Dataset Preparation:**
   - The Ego4D dataset was pre-processed to extract relevant video clips and their corresponding natural language queries.
   - Omnivore and EgoVLP features were extracted from these clips to be used as inputs to the models.
2. **Feature Integration:**
   - **Omnivore Features:** General visual capabilities captured from the video clips were directly fed into the models.
   - **EgoVLP Features:** These features, which are designed for egocentric vision tasks, were utilized to evaluate their impact on model performance.
3. **Feature Encoder Configuration:**
   - Two configurations were tested for the VSLNet model with Omnivore and EgoVLP features:
     - **Shared Encoder** Both the video and text encoders had a kernel size of 7.
     - **Non-Shared Encoder:** The video encoder's kernel size was increased to 9, while the text encoder's kernel size remained at 7. This allowed the video encoder to cover broader aspects of the input videos and extract more complex features while the text encoder remained focused on key parts of the textual input.
4. **Model Configuration:**
   - **VSLBase:** A simplified version of VSLNet, providing a point of comparison against the enhanced VSLNet.
   - **VSLNet:** Enhanced with a Query Guided Highlighter module to improve the precise temporal alignment between video and textual inputs.
5. **Training and Validation:**
   - Each model-feature combination was extensively trained with similar hyperparameters to ensure a fair comparison.
   - Performance metrics were evaluated on a validation set using Top-k Recall and mIoU (mean Intersection over Union) metrics to assess spatial-temporal accuracy.

4.2. Results

The results of the experiments for the different configurations are summarized in Table 1. These results show the performance of each model-feature combination on the NLQ benchmark across different evaluation metrics.

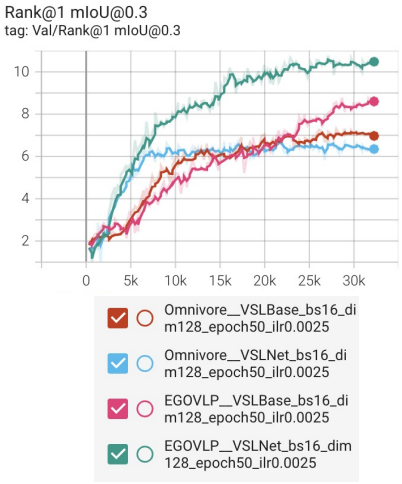

Figure 1. Comparison of Omnivore VSLBase vs. Omnivore VSLNet vs. EgoVLP VSLBase vs. EgoVLP VSLNet

4.3. Analysis

1. **Performance Comparison:**
   - **Omnivore + VSLBase:** This combination showed the lowest performance among all configurations. The overall mIoU of 4.52 indicates that while the model could identify relevant segments, its precision was relatively lower

Table 1. Performance Metrics for Different Model-Feature Configurations

Configuration	Rank@1 mIoU@0.3	Rank@1 mIoU@0.5	Rank@1 mIoU@0.01	Rank@3 mIoU@0.3	Rank@3 mIoU@0.5	Rank@3 mIoU@0.01	Rank@5 mIoU@0.3	Rank@5 mIoU@0.5	Rank@5 mIoU@0.01	mIoU
Omnivore + VSLBase	6.01	3.18	17.66	9.99	6.01	29.43	12.24	7.69	36.11	4.52
EGOVLP + VSLBase	8.65	5.24	**25.68**	13.89	9.11	**41.25**	16.68	11.18	**48.99**	6.67
EGOVLP + VSLNet (Shared Encoder)	**10.51**	5.81	23.62	**16.03**	9.73	37.22	**19.13**	**12.21**	45.10	**7.63**
EGOVLP + VSLNet (Non-Shared Encoder)	10.14	**6.09**	23.36	15.02	**9.99**	38.75	17.58	12.16	46.67	7.34
Omnivore + VSLNet (Shared Encoder)	6.38	3.51	15.75	10.22	6.12	27.23	11.85	7.54	34.36	4.76
Omnivore + VSLNet (Non-Shared Encoder)	6.38	3.61	15.07	9.78	5.94	26.95	11.72	7.41	33.61	4.77

Table 2. Performance of the NLQ baselines on val and Our EGOVLP + VSLNet (Shared Encoder)

Models	Rank@1 IoU=0.3 (%)	Rank@1 IoU=0.5 (%)	Rank@5 IoU=0.3 (%)	Rank@5 IoU=0.5 (%)
2D-TAN (Baseline)	5.04	2.02	12.89	5.88
VSLNet (Baseline)	5.45	3.12	10.74	6.63
EGOVLP + VSLNet (Shared Encoder)	**10.51**	**5.81**	**19.13**	**12.21**

compared to configurations using EgoVLP features or the more advanced VSLNet model. It achieved Rank@1 mIoU@0.5 of 3.18 and Rank@1 mIoU@0.01 of 17.66.

- **EgoVLP + VSLBase:** Pairing the VSLBase model with EgoVLP features significantly improved performance, highlighting the effectiveness of these specialized features in capturing egocentric vision details. This configuration outperformed the Omnivore + VSLBase across all metrics, demonstrating the benefit of using features pre-trained on egocentric video-language tasks. It achieved Rank@1 mIoU@0.5 of 5.24 and overall mIoU of 6.67.
- **EgoVLP + VSLNet (Shared Encoder):** The shared encoder configuration of VSLNet with EgoVLP features yielded strong performance, achieving the highest overall mIoU of 7.63. This suggests that using the same kernel size for both video and text encoders facilitates better integration and alignment between video and text features, leading to more accurate temporal localization. It achieved Rank@1 mIoU@0.5 of 5.81 and Rank@1 mIoU@0.01 of 23.62.
- **EgoVLP + VSLNet (Non-Shared Encoder):** The non-shared encoder configuration also performed well but slightly less than the shared encoder version. An overall mIoU of 7.34 indicates that while increasing the kernel size of the video encoder captured more extensive video context, it added complexity that slightly hindered the precise alignment of video and text features. It achieved Rank@1 mIoU@0.5 of 6.09 and Rank@1 mIoU@0.01 of 23.36.
  - **Explanation**: The plot in Figure 2 shows the comparison between EgoVLP + VSLNet with shared and non-shared encoders. The shared encoder configuration achieved a slightly higher Rank@1 mIoU@0.3 compared to the non-shared encoder, indicating that using the same kernel size for both video and text encoders facilitated better integration and improved temporal localization performance.

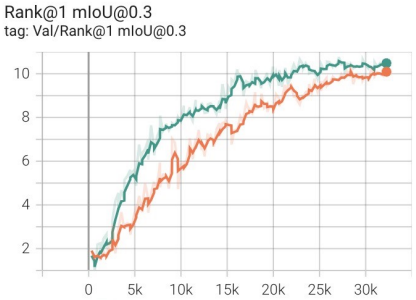

Figure 2. Comparison of EgoVLP VSLNet Shared Encoder vs. Non-Shared Encoder

- **Omnivore + VSLNet (Shared Encoder):** Using the same kernel size (7) for both video and text encoders resulted in a balanced performance, suggesting a balanced feature extraction process. This configuration achieved Rank@1 mIoU@0.5 of 3.51 and overall mIoU of 4.76.
- **Omnivore + VSLNet (Non-Shared Encoder):** Increasing the kernel size of the video encoder to 9 while keeping the text encoder's size at 7 resulted in slightly lower performance. This configuration achieved Rank@1 mIoU@0.5 of 3.61 and overall mIoU of 4.77, indicating that while a larger kernel size for the video encoder allowed it to capture more extensive video context, the resulting complexity may have slightly hindered

the model's ability to precisely align the video and text features.

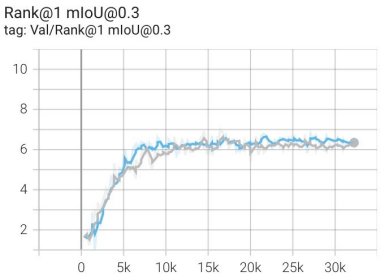

Figure 3. Comparison of Omnivore VSLNet Shared Encoder vs. Non-Shared Encoder

  - **Explanation**: The plot in Figure 3 compares Omnivore + VSLNet with shared and non-shared encoders. The performance difference between shared and non-shared encoders is minimal, but the shared encoder configuration slightly outperforms the non-shared encoder in terms of Rank@1 mIoU@0.3, indicating better integration and alignment of video and text features.

2. **Impact of Feature Encoder Configuration:**

   - **Shared Encoder (Kernel Size 7 for Both):** When the encoders shared the same kernel size, the performance was slightly better in terms of overall mIoU compared to the non-shared encoder for Omnivore features, showing a more straightforward integration between video and text features.
   - **Non-Shared Encoder (Kernel Size 9 for Video, 7 for Text):** While this approach aimed to capture more complex video features, the increase in kernel size for the video encoder only provided a marginal improvement or slight decrease in performance compared to the shared encoder configuration.

3. **Metric Insights:**

   - **Rank@1, Rank@3, Rank@5:** These recall metrics demonstrate the models' ability to correctly predict the relevant temporal segments within the top k predictions. Higher values indicate better performance, with EgoVLP + VSLNet configurations generally performing better across these metrics.
   - **mIoU@0.3, mIoU@0.5, mIoU@0.01:** These mIoU metrics at different thresholds highlight how accurately the models can align predicted segments with the ground truth. Higher values suggest superior temporal precision, with EgoVLP + VSLNet showing the highest precision.
   - **Overall mIoU:** The mean IoU provides a comprehensive measure of the overlap between predicted segments and the ground truth, reflecting overall alignment accuracy. EgoVLP + VSLNet (Shared Encoder) had the highest overall mIoU at 7.63, indicating the best performance among the configurations.

4.4. Discussion

The experimental results provide several insights into how model architecture and feature selection impact temporal localization tasks:

1. **Model and Feature Synergy:**

   - The advanced VSLNet model paired with EgoVLP features consistently outperformed other combinations, underscoring the importance of using specialized features and sophisticated models for complex tasks like egocentric vision. Compared to the baseline provided in the Ego4D paper, our best configuration (EgoVLP + VSLNet with shared encoder) showed a significant improvement in mIoU, demonstrating the effectiveness of the combined features and advanced model architecture.

2. **Encoder Configuration:**

   - The slight performance drop when using non-shared encoders (different kernel sizes for video and text) suggests that the added complexity from a larger video kernel size may not always translate to better performance. Shared configurations can sometimes yield more robust results, particularly in terms of integration and alignment between different input modalities.

3. **Future Directions:**

   - Further exploration of different feature sets, especially those designed for specific aspects of egocentric vision, could potentially enhance performance.
   - Experimenting with more advanced or varied encoder configurations could reveal additional insights into optimizing model architecture for temporal localization tasks.

- Investigating additional enhancements to the VSLNet model or exploring other model architectures that leverage advanced attention mechanisms could further improve performance.

5. Conclusion

The project demonstrated the critical importance of specialized features and sophisticated models for addressing the Natural Language Queries (NLQs) task in egocentric vision. The combination of EgoVLP features with the VSLNet model, particularly in the shared encoder configuration, achieved the highest performance, highlighting the effectiveness of integrated and aligned video-text features. Compared to the baseline results from the Ego4D paper, our approach showed significant improvements, underscoring the potential of using advanced feature sets and model architectures for complex video understanding tasks. This research provides valuable insights for future exploration in the domain of egocentric vision, with potential applications in personal assistance, accessibility, behavioral analysis, and education. The extension using Large Language Models (LLMs) further enhances interaction with egocentric video data, showcasing the potential of advanced language models in generating natural language queries. Further exploration and optimization in this area could lead to even more robust and accurate systems for understanding and interacting with egocentric videos.

6. Extension

We implement a natural language query (NLQ) generation system in this extension using Google's open-source model, Gemma 2B EN (ar5iv). This system leverages dependencies installed via Kaggle's API using a provided username and API key. For the backend framework, we have chosen JAX due to its efficiency and flexibility in handling large-scale model training.

To achieve our desired results, we train the Gemma 2B model with our specific data, ensuring the output aligns with the required format. This involves generating a comprehensive dataset of input-output pairs using the nlq_train.json file. This file contains two properties that serve as our input and output formats:

- **Input (descriptive sentence):** slot_y or slot_x
- **Output (question about the input):** query

By utilizing these properties, we create a large dataset where each input is a descriptive sentence about an object or person, and each output is a corresponding question about that descriptive sentence.

For fine-tuning the model, we employ the LoRA (Low-Rank Adaptation) technique introduced by Google (ar5iv). This method, known for its efficiency in adapting large language models to specific tasks with relatively small computational overhead, reassures the model's flexibility.

The training process on an A1000 GPU in Google Colab Pro takes approximately 45 minutes. After training, the model can generate questions based on input descriptive sentences, facilitating intuitive querying and interaction with egocentric video data.

References

[1] K. Grauman *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022. 1, 2

[2] D. Damen *et al.*, "Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100," *Int. J. Comput. Vis.*, 2022. 1

[3] J. Gao *et al.*, "Tall: Temporal activity localization via language query," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017. 1, 2, 3

[4] H. Zhang *et al.*, "Span-based Localizing Network for Natural Language Video Localization," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020. 2, 3

[5] C. Feichtenhofer *et al.*, "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019. 2

[6] K. Q. Lin *et al.*, "Egocentric video-language pretraining," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 7575-7586, 2022. 1, 2, 3

[7] S. Zhang *et al.*, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020. 1

[8] J. Van Amersfoort, A. Kannan, M. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-based models of video sequences," *arXiv preprint arXiv:1701.08435*, 2017. 3

[9] H. Le, D. Sahoo, N. Chen, and S. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, pp. 5612–5623, 2019.

[10] S. Ghosh, A. Agarwal, Z. Parekh, and A. Hauptmann, "ExCL: Extractive Clip Localization Using Natural Language Descriptions," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, pp. 1984–1990, 2019. 2, 3

[11] R. Girdhar *et al.*, "Omnivore: A single model for many visual modalities," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022. 2, 3

[12] B. Lin *et al.*, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023. 3