# Bridging Vision and Language: NLQ with Textual Answer Generation

Leone Fabio, Gabriele Raffaele, Giuseppe Vacante
Politecnico Di Torino
s330500@studenti.polito.it, s331156@studenti.polito.it, s332155@studenti.polito.it

## Abstract

*This paper explores the Ego4D dataset and its Natural Language Queries (NLQ) annotations, proposing a novel pipeline that integrates timestamp prediction models with a Video Question Answering (VideoQA) module. The pipeline transitions from untrimmed egocentric video inputs to actionable textual answers, addressing challenges in computational efficiency and detail preservation inherent in video analysis. We benchmark state-of-the-art models—VSLBase, VSLNet and 2D-TAN—trained on Omnivore and EgoVLP features to evaluate their performance in temporal segment localization. Building on these insights, our approach refines timestamp predictions into precise video segments processed by the videoQA model to generate textual answers. Experimental results demonstrate improvements in temporal localization and question answering, showcasing the integration of video reasoning and natural language understanding. This work advances egocentric video analysis, with applications in episodic memory, assistive technology, and automated reasoning. Code, results and statistics are publicly available at https://github.com/LeoneFabio/Egocentric-Vision.*

## 1. Introduction

Egocentric video analysis, exemplified by the Ego4D dataset [17], captures complex human activities but presents challenges due to untrimmed, unstructured content. The NLQ benchmark [17] identifies temporal intervals, answering a query, but requires users to manually watch these segments, limiting usability. To address this, we propose a pipeline combining timestamp prediction models with the Video-LLaVA [26] VideoQA module. This approach transforms unstructured videos into precise textual answers, optimizing computational efficiency while preserving details. Benchmark analysis of VSLBase, VSLNet [20], and 2D-TAN [41] models trained with Omnivore [15] and EgoVLP [22] features highlights their effectiveness in temporal localization. Video-LLaVA [26] extends these results by processing trimmed segments into actionable answers, reducing overhead and enhancing usability. Our contributions include:

1. Benchmark analysis of NLQ models with advanced features.
2. Integration of timestamp prediction with VideoQA.
3. Validation showing improved performance for egocentric video analysis systems.

These advancements demonstrate the potential of our approach for episodic memory retrieval, assistive technologies, and video understanding.

## 2. Related Work

**Natural Language Queries in Egocentric Videos.** The Natural Language Queries (NLQ) task involves localizing the temporal window corresponding to the answer to a question in a long video clip. This task is challenging for end-to-end supervised video localization models due to the sparsity of annotations and the length of videos in the dataset. However, strong video representations, such as those derived from SlowFast [12], Omnivore [15], and EgoVLP [22], significantly simplify the task. Several notable models have been applied to the NLQ benchmark. VSLNet [20], a video span localization network, employs a query-guided proposal mechanism to effectively align temporal spans with natural language queries. 2D-TAN [41] extends temporal action detection to a two-dimensional framework, enabling efficient proposal generation and ranking for temporal localization. These models have been instrumental in advancing temporal localization and serve as baseline architectures in this study. Prior works have focused on constructing a hierarchical structure, augmenting the NLQ dataset and developing better video features through large-scale pre-training. ReLER [29] proposes a novel multi-scale cross-modal transformer architecture, a video frame-level contrastive loss, and two data augmentation strategies. InternVideo [19] improves the quality of video features by carefully pre-training and fine-tuning a VideoMAE-L Model [40], and ensemble the features and predictions. More recently, NaQ [30] introduces a data augmentation strategy

to transform video narrations into training data for the NLQ task, alleviating the problem of sparse annotation. NaQ++ ReLER, obtained by training the ReLER model with NaQ data, was the previous state-of-the-art method for Ego4D NLQ. GroundNLQ [42] is the current state-of-the-art for this benchmark. It adopts a two-stage pre-training strategy to respectively train a video feature extractor and a grounding model on video narrations, and finally finetune the grounding model on annotated data. Our work is complementary to these prior efforts, as they can be leveraged in the first stage of our proposed framework to localize temporal segments, which are subsequently refined into fine-grained textual answers using a frozen VideoQA model. This approach bridges the gap between timestamp predictions and actionable insights, advancing the utility of NLQ models in practical applications.

**VideoQA**. Video Question-Answering (videoQA) is a key task for multimodal video understanding systems to assess their ability to reason about a video [18, 28, 35, 37, 38]. Recent benchmarks have pushed towards assessing reasoning for temporal questions [18, 35, 37], longer videos [28, 39], and on domains like instructional [38] and egocentric videos [28, 39].

**End-to-end Models for VideoQA.** The recent success of LLMs [2, 7, 16, 32] has led to an explosion of multimodal models that jointly understand vision and text data. Many works map frozen image encoders [9, 10, 35] to the LLM textual embedding space: e.g., Flamingo [1], via a Perceiver resampler [21], or BLIP2 [25] and Video-LLaMa [24], via Q-formers for audio/vision [10, 14]. GIT2 [33] and PALI [2, 4, 5] use simple encoder-decoder style architectures which are trained for image captioning, while MV-GPT [31] finetunes a native video backbone [11] for video captioning. Although trained with a generative (captioning) objective, such models achieve strong results for general vision-language tasks (cast as auto-regressive generation with question as prefix). More recent works such as InstructBLIP [8], MiniGPT-4 [43], and VideoBLIP [3] improve zero-shot results with strong instruction tuning. Generally, however, end-to-end methods can be difficult to interpret. For videos in particular, memory limits in end-to-end models require significant downsampling: e.g., temporally sampling a few frames with large strides [2, 33], spatially subsampling each frame to a single token [23, 27, 34]. Such models also tend to process each frame with equal importance. Unlike such works, our model has an explicit grounding stage, which searches for the most relevant video frames to be processed in more detail. Other grounding works for videoQA include SeViLa [6], MIST [13], and NExT-GQA [36].

However, these end-to-end models often require significant downsampling and struggle to process long videos efficiently, making them less suitable for high-detail video reasoning tasks. To address these limitations, our work adopts a modular approach, where the temporal localization (NLQ) model provides fine-grained segment predictions, and **Video-LLaVA** is then used to generate textual answers from those segments. This two-stage pipeline enhances the model's efficiency and answer quality.

## 3. Methodology

This section outlines the steps taken to analyze the training dataset, train and evaluate temporal localization models, and extend the pipeline for generating textual answers from localized video segments.

### 3.1. Dataset and Annotation Analysis

We employed the Ego4D dataset, focusing on its Natural Language Queries (NLQ) annotations. This dataset comprises approximately 19,000 queries from 227 hours of egocentric video content, with each query annotated with temporal boundaries indicating when the query is answered. To better understand the Ego4D training dataset and its Natural Language Queries (NLQ) annotations, we conducted an extensive analysis focusing on template distributions, clip and answer segment durations, temporal relationships and observations on scenarios. These insights informed the design and evaluation of our models and ensured they accounted for the dataset's diverse characteristics.

| Statistic | Value (seconds) |
|---|---|
| Average Clip duration | 522.68 |
| Max Clip duration | 1200.07 |
| Min Clip duration | 207.17 |
| Standard deviation | 197.64 |
| Median Clip duration | 480.00 |

Table 1. Summary statistics for clip durations, including mean, median, standard deviation, minimum, and maximum values, highlighting the variability in clip lengths within the dataset.

The distribution of query templates, visualized in Figure 1a, revealed the presence of queries with the "None" template, where no specific template was assigned. The histogram in Figure 1b, complemented by statistical metrics in Table 1, illustrates the wide range of clip durations, which span from approximately 207 seconds to 20 minutes, with a mode at 480 seconds. This variation underscores the necessity for temporal localization models to handle diverse input lengths effectively.

The answer segment durations, visualized in Figure 1c and detailed through additional statistics in Table 2, showed

| Statistic | Value (seconds) |
|---|---|
| Average Answer segment duration | 9.67 |
| Max Answer segment duration | 480.00 |
| Min Answer segment duration | 0.00 |
| Standard deviation | 22.83 |
| Median Answer segment duration | 3.45 |

Table 2. Summary statistics for answer segment durations, showcasing key metrics such as mean, median, standard deviation, minimum, and maximum values, reflecting the predominantly short length of answer segments in the dataset.

that these segments are generally short, ranging from 0 to 480 seconds, with a median of 3.45 seconds. Notably, the presence of zero-second durations prompted us to apply a filter to VSLNet, VSLBase and 2D-TAN model to exclude such no-time answer segments, enhancing the robustness of the models.

Finally, additional statistics on query counts and answer durations across scenarios highlight significant variability in these attributes, as shown in Table 3 and Table 4. This variability emphasizes the dataset's complexity and reinforces the importance of scenario-specific considerations in model design.

| Statistic | Value |
|---|---|
| Average Query count | 223.99 |
| Max Query count | 3018 |
| Min Query count | 4 |
| Standard deviation | 455.08 |
| Median Query count | 73.0 |

Table 3. Statistics on query counts across scenarios, highlighting the variability in the number of queries per scenario with key metrics such as mean, median, standard deviation, minimum, and maximum values.

| Statistic | Value (seconds) |
|---|---|
| Average Answer duration | 10.69 |
| Max Answer duration | 54.58 |
| Min Answer duration | 1.32 |
| Standard deviation | 7.05 |
| Median Answer duration | 9.72 |

Table 4. Statistics on answer durations across scenarios, providing insights into the distribution of answer segment lengths with key metrics including average, maximum, minimum, standard deviation, and median values.

### 3.2. Temporal Localization Model Training

We trained, validated, and tested three temporal localization models—VSLBase, VSLNet and 2D-TAN—on pre-extracted features from Omnivore [15] and EgoVLP [22].

- VSLBase [20], served as a foundational architecture, extracting visual and textual features that are fused through shared encoders and refined with Context-Query Attention. Temporal boundaries were regressed using LSTMs.
- VSLNet [20], extending VSLBase, incorporated a Query-Guided Highlighter for finer temporal alignment, improving its ability to handle subtle differences in video frames.
- 2D-TAN [41], framed temporal localization as a two-dimensional proposal generation and ranking process, emphasizing complementary strengths to VSLNet.

Training was performed with official pre-extracted features, allowing computationally efficient fine-tuning of these architectures.

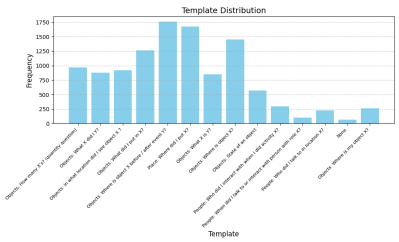
### 3.3. Extension to Video Question Answering

To bridge the gap between temporal localization and textual answer generation, we extended the NLQ task using a VideoQA model, Video-LLaVA [26]. Our approach involved the following steps:

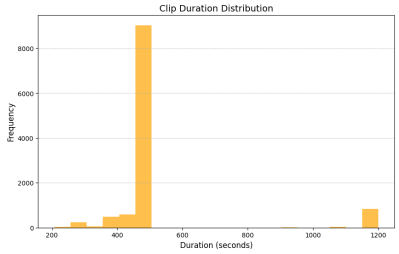
1. **Top Predictions Selection**: We selected the top 50 VSLNet predictions based on Intersection over Union (IoU) scores, ensuring high-quality temporal segments.
2. **Video Segments Extraction**: Using the predictions, the corresponding video segments were extracted with ffmpeg.
3. **VideoQA Model Inference**: Each segment, paired with its query, was fed into Video-LLaVA to generate textual answers.
4. **Evaluation**: The generated answers were assessed using a comprehensive set of metrics, including SACREBLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, and METEOR, providing a nuanced evaluation of relevance, fluency, and semantic alignment.

## 4. Experiments

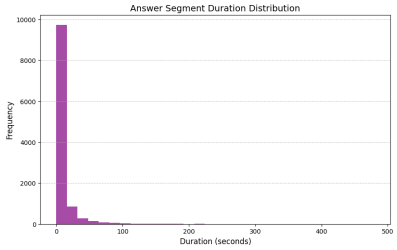
This section presents the experiments carried out to evaluate the effectiveness of the temporal localization models and the VideoQA extension. We examine the performance of VSLBase, VSLNet and 2D-TAN models trained on different feature sets (Omnivore, and EgoVLP) and compare their results to the baseline models trained on SlowFast features from the Ego4D research.



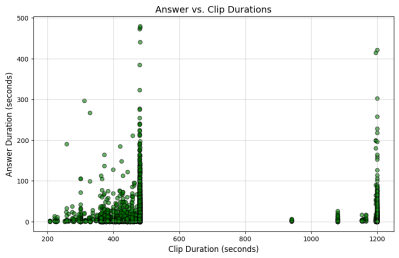

(a) **Template Distribution**: Bar chart of query frequencies by template, including "None."



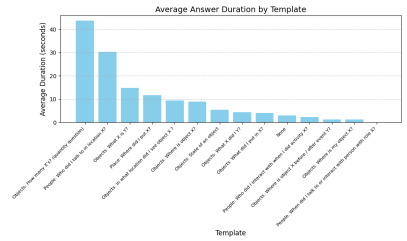

(b) **Clip Duration Distribution**: Histogram of input clip durations, ranging from 207 seconds to 20 minutes, with a peak at 480 seconds.



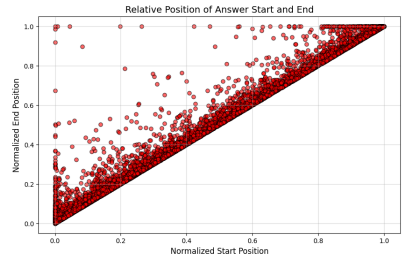

(c) **Answer Segment Duration Distribution**: Histogram of ground truth answer durations, mostly short (0–480 seconds), with zero-duration answers notable.



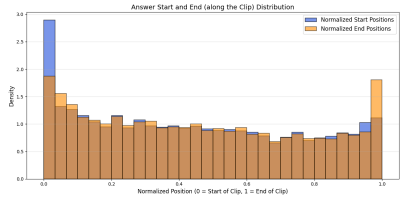

(d) **Answer vs. Clip Durations**: Scatter plot showing the relationship between clip and answer durations.




(e) **Average Answer Duration by Template**: Bar chart showing quantity-based queries have the longest average answers.




(f) **Relative Position of Answer Start and End**: Scatter plot of normalized start and end positions of answers relative to clips.




(g) **Answer Start and End Distribution**: Overlapping histograms of normalized start and end timestamps along clips.

Figure 1. Visualization of key dataset characteristics and distributions.

### 4.1. Temporal Localization

To evaluate the performance of the temporal localization models, we trained and validated VSLBase, VSLNet and 2D-TAN using the Omnivore and EgoVLP features. Results were also compared to the Ego4D baselines trained on SlowFast features (as presented in Table 7). The evaluation metrics primarily consisted of **Intersection over Union** (IoU) at different thresholds. The results, shown in Table 5 and Table 6, offer key insights into the influence of different features and model architectures.

**Settings.** To ensure fair and consistent evaluation of the models in a reasonable time, specific hyperparameter settings were adopted during training. For **VSLBase** and **VSLNet**, the batch size was set to 32, with the embedding dimension (DIM) defined as 128. Training was conducted over 10 epochs, and the maximum position length was configured to 128. Additionally, an initial learning rate of 0.0025 was used to facilitate effective optimization. In the case of **2D-TAN**, the hyperparameters were similarly tailored to the model's requirements. A batch size of 32 was employed, with a learning rate of 0.025 and no weight decay applied. The training process spanned a maximum of 5 epochs. For video segmentation, 40 sample clips were used, with a target stride of 1. The data was normalized and subjected to random sampling to enhance robustness. Furthermore, a 40-second window size was utilized to segment the videos effectively. To manage training and validation within a limited timeframe due to resource constraints, only 1/10th of the training dataset and 1/16th of the validation dataset of NLQ annotations were used for **2D-TAN**. These hyperparameter settings were carefully selected based on insights from prior research and iterative fine-tuning experiments.

**Performance on Omnivore Features** (Table 5): When

| Model | Rank@1, IoU@0.3 | Rank@1, IoU@0.5 | Rank@5, IoU@0.3 | Rank@5, IoU@0.5 |
|---|---|---|---|---|
| VSLBase | 6.14 | 3.46 | 12.65 | 7.90 |
| VSLNet | 6.56 | 3.95 | 13.16 | 8.05 |
| 2D-TAN | 10.53 | 2.63 | 13.16 | 10.53 |

Table 5. Results for VSLBase, VSLNet, and 2D-TAN on Omnivore Features

| Model | Rank@1, IoU@0.3 | Rank@1, IoU@0.5 | Rank@5, IoU@0.3 | Rank@5, IoU@0.5 |
|---|---|---|---|---|
| VSLBase | 5.76 | 3.41 | 12.52 | 7.92 |
| VSLNet | 6.43 | 4.05 | 13.86 | 9.22 |
| 2D-TAN | 8.51 | 3.55 | 17.02 | 7.80 |

Table 6. Results for VSLBase, VSLNet, and 2D-TAN on EgoVLP Features

| Model | Rank@1, IoU@0.3 | Rank@1, IoU@0.5 | Rank@5, IoU@0.3 | Rank@5, IoU@0.5 |
|---|---|---|---|---|
| 2D-TAN | 5.04 | 2.02 | 12.89 | 5.88 |
| VSLNet | 5.45 | 3.12 | 10.74 | 6.63 |

Table 7. Results for 2D-TAN and VSLNet on SlowFast Features (Ego4D Baseline)

the models were trained on Omnivore features, the results show that **VSLNet** outperforms **VSLBase** across all metrics. For example, VSLNet achieved a **Rank@1** of 6.56 at IoU@0.3, compared to VSLBase's 6.14, suggesting that VSLNet is more capable of precise temporal localization when trained on Omnivore features. Furthermore, **2D-TAN**, while showing slightly lower performance in Rank@1 metrics, exhibits competitive results, particularly at Rank@5. This highlights the strength of 2D-TAN's proposal-based approach, which complements VSLNet in handling more general queries that require broader temporal context.

**Performance on EgoVLP Features** (Table 6): The models trained on **EgoVLP features** demonstrate a noticeable improvement in performance, particularly in the case of VSLNet. For instance, VSLNet achieves generally higher results than its performance on Omnivore features. This suggests that EgoVLP, which integrates more robust video-language representations, improves the model's ability to localize answers more accurately within video segments. Additionally, the **2D-TAN** model shows a a different trend, with some values that are lower than them counterpart on Omnivore.

**Comparison with SlowFast Baselines** (Table 7): When compared to the **SlowFast** features used in the original Ego4D paper, the models trained on Omnivore and EgoVLP features show considerable advantages in performance. Specifically, **VSLNet** on SlowFast achieves a **Rank@1** of 5.45 at IoU@0.3, which is lower than the performance observed on both Omnivore (6.56) and EgoVLP (6.43) features. Similarly, **2D-TAN** outperforms the SlowFast baseline in both Rank@1 and IoU at multiple thresholds, indicating that the additional information provided by Omnivore and EgoVLP features improves the models' ability to localize temporal segments more effectively.

#### 4.1.1 Observations

**Model Comparison**: Across all feature sets, **VSLNet** consistently outperforms **VSLBase**, confirming its superiority in capturing finer temporal details and understanding the relationship between queries and video content. The integration of a Query-Guided Highlighter in VSLNet likely enhances its sensitivity to subtle changes in the video's temporal structure, which is reflected in its higher precision scores.

**Role of Feature Sets**: The results highlight the importance of the input features in the performance of temporal localization models. **EgoVLP** and **Omnivore** features, which provide richer multimodal representations, allow the models to more effectively leverage both visual and textual information. The models trained on these features consistently outperform the baselines using **SlowFast** features, which are more limited in their ability to represent complex video-language interactions.

**Proposal-Based Strengths of 2D-TAN**: While **2D-TAN** lags slightly behind VSLNet in certain metrics, it shows complementary strengths in handling queries with broader temporal contexts, where more generalized proposals are required. Its performance in Rank@5 and IoU scores indi-

cates that it excels at localizing less precise but still important segments within the video.

In conclusion, the experiments demonstrate that both VSLNet and 2D-TAN benefit significantly from richer video language representations, such as those provided by EgoVLP and Omnivore features, outperforming the baselines trained on SlowFast features. These results underline the importance of feature selection and model architecture in the task of temporal localization, with implications for improving both the accuracy and efficiency of video understanding systems.

### 4.2. Video Question Answering

Using the top 50 VSLNet predictions, we extended the NLQ task to VideoQA by leveraging the Video-LLaVA model. The preprocessing for the extracted video in Video-LLaVA involves two steps: (1) uniformly sampling 8 frames from each video by dividing it into 8 segments, and (2) converting each selected frame into a 24-bit RGB array for further processing. The pipeline processed the frames and generated textual answers aligned with the input queries.

- **Qualitative Results**: The generated answers demonstrated relevance and coherence with the input queries. Selected examples [A] illustrated the effectiveness of localizing and processing critical video segments.
- **Quantitative Results**: To assess performance, the generated answers were compared against ground-truth annotations using a comprehensive set of evaluation metrics: SACREBLEU, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, and METEOR.
  - **SACREBLEU** scores ranged between (1.93% - 59.46%) with a mean of 19.29%.
  - **ROUGE-1, ROUGE-2, ROUGE-L**, and **METEOR** scores demonstrated consistent alignment with ground-truth annotations, with values centered around 0.49, 0.27, 0.45, 0.48 respectively.
  - **BERTScore** reflected strong semantic similarity between generated answers and ground truth, underscoring the model's capacity for nuanced understanding, earning an average score of 0.92.

Table 8 consolidates the additional statistics for all metrics, providing a comprehensive summary of the evaluation results. Figure 2 presents the scatter plots for each evaluation metric, where each plot visualizes the distribution of scores for the generated answers. Figure 3 shows the histograms of the evaluation metrics, illustrating the frequency distribution of the scores for each metric.

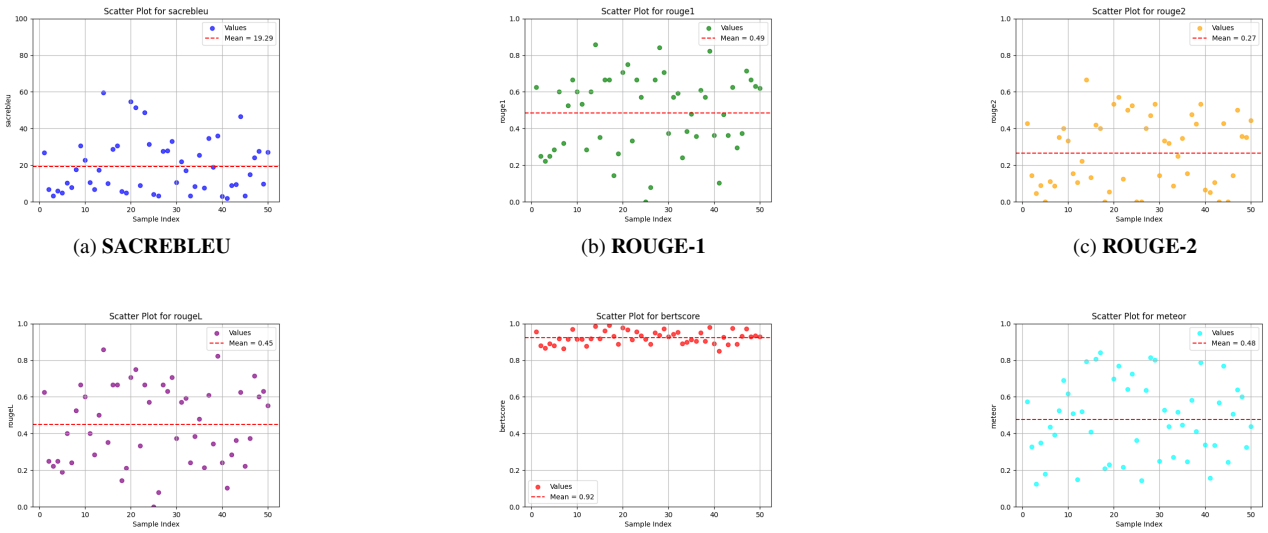
| Metric | Mean | Min | Max | Median | Std Dev |
|---|---|---|---|---|---|
| SacreBLEU (%) | 19.29% | 1.93% | 59.46% | 15.99% | 15.10% |
| Rouge1 | 0.49 | 0.00 | 0.86 | 0.55 | 0.21 |
| Rouge2 | 0.27 | 0.00 | 0.67 | 0.29 | 0.19 |
| RougeL | 0.45 | 0.00 | 0.86 | 0.44 | 0.21 |
| BERTScore | 0.92 | 0.85 | 0.99 | 0.92 | 0.04 |
| METEOR | 0.48 | 0.12 | 0.84 | 0.48 | 0.21 |

Table 8. Summary statistics for all metrics

## 5. Conclusion

This work investigated the NLQ benchmark within the Ego4D dataset, focusing on temporal localization of answers and extending the task to textual answer generation. Through the use of VSLBase, VSLNet and 2D-TAN models with Omnivore and EgoVLP features, we demonstrated the importance of pre-trained representations and advanced architectures for accurate temporal localization. The extension to VideoQA, using top 50 VSLNet predictions and the Video-LLaVA model, showcased the feasibility of generating meaningful textual answers from localized segments.

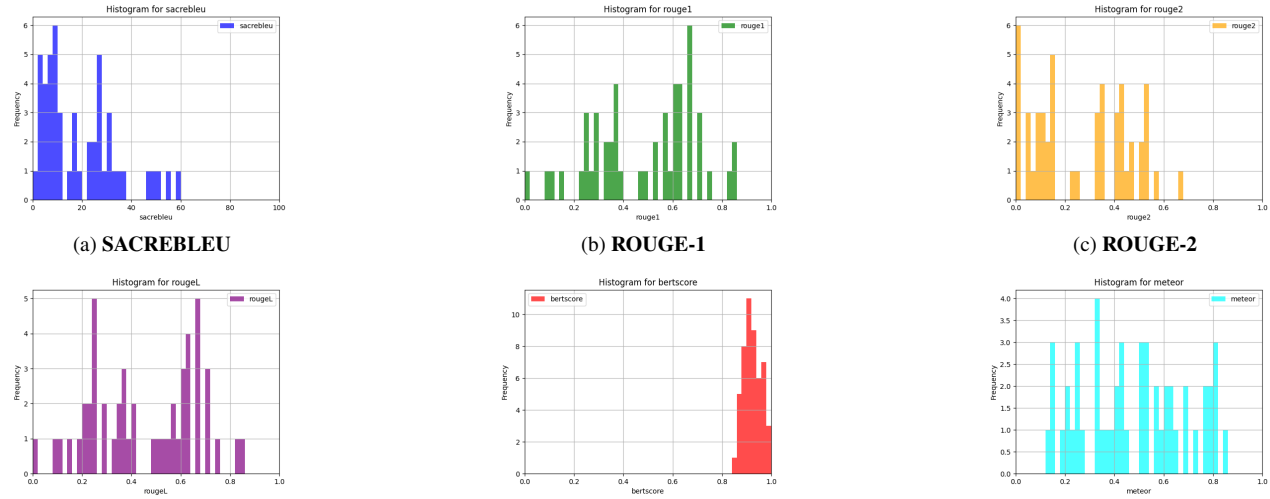
These results underscore the potential of integrating temporal localization and VideoQA, addressing the challenges of egocentric video analysis and opening avenues for real-world applications in assistive and automated systems.




(a) **SACREBLEU** (b) **ROUGE-1** (c) **ROUGE-2**

(d) **ROUGE-L** (e) **BERTSCORE** (f) **METEOR**

Figure 2. A red line indicates the mean value for each metric, offering a clear visual representation of the performance trends across the dataset.




(a) **SACREBLEU** (b) **ROUGE-1** (c) **ROUGE-2**

(d) **ROUGE-L** (e) **BERTSCORE** (f) **METEOR**

Figure 3. Presents the histograms depicting the distribution of values for each evaluation metric.


## Acknowledgement

Thanks to the Ego4D consortium for dataset access and support.


## References


[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, and et al. Pali-x: On scaling up a multilingual vision and language model. In *arXiv preprint arXiv:2305.18565*, 2023. 2

[3] Xi Chen, Junnan Li, Xiao Wang, Jean-Baptiste Alayrac, and Rohit Girdhar. Videoblip: Advancing video-language models with instructional training. *arXiv preprint arXiv:2307.02345*, 2023. 2

[4] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, and et

al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 2

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and et al. Pali: A jointly-scaled multilingual language-image model. In *arXiv preprint arXiv:2209.06794*, 2022. 2

[6] Xi Chen, Xiao Wang, Rohit Girdhar, Daniel Zhou, and Yuxin Fang. Sevila: A novel framework for visual-language grounding in videos. *arXiv preprint arXiv:2309.00239*, 2023. 2

[7] Aakanksha Chowdhery, Sharan Naranga, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[8] Wenliang Dai, Junnan Li, Dongxu Li, Yinan Yu, Weiying Wang, Siddharth Bhargava, Chaoyi Zhang, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[10] Yuxin Fang, Wen Wang, Binhui Xie, Xintao Zhang, Zhenda Zhang, Xiaoyu Tao, Xiangyang Ji, and Li Zhang. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[11] Christoph Feichtenhofer. Masked autoencoders for video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 774–784, 2022. 2

[12] Christoph Feichtenhofer and et al. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1

[13] Difei Gao, Luowei Zhou, Lei Ji, Rowan Zellers, Yi Yu, Bo Dai, Dongxu Li, Rohit Girdhar, Jianfeng Gao, Kristen Grauman, et al. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783, 2023. 2

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Andrew Rouditchenko, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[15] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 1, 3

[16] Google, Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, and et al. Palm 2 technical report. *arXiv preprint arXiv:2302.13971*, 2023. 2

[17] Kristen Grauman et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[18] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 2

[19] Chen Guozhong et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges, 2022. arXiv preprint arXiv:2211.09529. 1

[20] Zhang Hao and et al. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 1, 3

[21] Andrew Jaegle and Others. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021. 2

[22] Lin Kevin Qinghong and et al. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pages 7575–7586, 2022. 1, 3

[23] Jie Lei and Others. Less is more: Sampling strategies for efficient video processing. *arXiv preprint arXiv:2107.01360*, 2021. 2

[24] Dongxu Li, Junnan Li, Chaoyi Zhang, Wenliang Dai, and Xiao Wang. Video-llama: Towards video-language models with expanded contexts. *arXiv preprint arXiv:2308.09199*, 2023. 2

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[26] Bingqian Lin, Bowen Zhu, Yixiao Ye, Mengshi Ning, Peng Jin, and Lu Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 3

[27] Wenhao Liu and Others. Optimize frame sampling for vision tasks. *arXiv preprint arXiv:2203.04565*, 2022. 2

[28] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023. 2

[29] Liu Ning, Wang Xiaohan, Li Xinyi, Yang Yujing, and Zhuang Yueting. Reler@zju-alibaba submission to the ego4d natural language queries challenge 2022, 2022. arXiv preprint arXiv:2207.00383. 1

[30] Shivansh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Naq: Leveraging narrations as queries to supervise episodic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[31] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF*

*Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 2

[32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[33] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Chen, Jianfeng Gao, Yejin Choi, William Hwang, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2

[34] Teng Wang, Ruimao Zhang, Zhichao Lu, Qingqiu Huang, Junliang Xing, Yan Gao, Zhihan Gao, and Haojie Hu. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021. 2

[35] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021. 2

[36] Li Xiao and Others. Next-gqa: Benchmarking video reasoning in open-ended domains. *arXiv preprint arXiv:2205.11038*, 2022. 2

[37] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of ACM Multimedia*, 2017. 2

[38] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 2

[39] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, , and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 2

[40] Tong Zhan, Song Yibing, Wang Jue, and Wang Limin. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 1

[41] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12770–12777. AAAI Press, 2020. 1, 3

[42] Hou Zhihan et al. Groundnlq @ ego4d natural language queries challenge 2023, 2023. arXiv preprint arXiv:2306.15255. 2

[43] Deyao Zhu and Others. Minigpt-4: Enhancing vision-language understanding with a smaller gpt. *arXiv preprint arXiv:2304.06590*, 2023. 2

## A. Example Appendix

This appendix provides representative examples of the input frames, the corresponding questions, and the generated answers from the Video-LLaVA pipeline. These examples illustrate how the system processes video segments and produces coherent and contextually relevant answers.

Example 1:


Where was the egg before I picked it?

The egg was in the refrigerator before you picked it

Example 2:


How many onions was chopped?

The person in the video chopped two onions.

Example 3:


What color is the paper I moved with my right hand?

The paper you moved with your right hand is white.