Machine Learning and Deep Learning project: Natural Language Queries for Egocentric Vision

Giovanni Andrea Brullo
s317663@studenti.polito.it

Abstract

This project explores the use of natural language queries for egocentric video understanding, leveraging the extensive Ego4D dataset and its NLQ benchmark task. Our work is divided into two phases: video moment localization and the generation of textual answers based on the identified moments. In the first phase, we address the NLQ task by identifying the temporal segment of the video where a given natural language query is answerable. We model this as a regression problem using video span localizing networks: VSLBase and its enhanced version, VSLNet. Both networks learn cross-modal interactions between video and text features through an attention-like mechanism, and they are able to regress the span of the target moment. By incorporating the Query-Guided Highlighting (QGH) module, VSLNet improves performance by focusing the search of the target moment within the highlighted region.We trained and validated this model using various combinations of powerful pre-trained models as fixed feature extractors for text and visual features on the NLQ benchmark task. In the second phase, we extend the project by extracting textual answers from the identified target moments using a large vision-language model, Video-LLaVA. This project demonstrates progress towards developing an augmented reality assistant capable of interpreting daily-life egocentric videos and responding to queries about past activities.

1. Introduction

Egocentric videos, captured by head-mounted cameras, provide a unique first-person perspective on human activities. These recordings document interactions and events from the individual's viewpoint, capturing the who, what, when, and where of their life experiences. This makes them particularly suitable for natural language queries on the cameraman's past experiences within the video content. The development of Ego4D, a massive egocentric video dataset featuring hundreds of daily life scenarios aligned with millions of annotations supporting complex tasks, provides the resources to tackle the NLQ benchmark task [1]. This task involves responding to a natural language query by identifying the temporal window within an untrimmed video where the answer is visible. The queries follow specific templates, as shown in Figure 1, and may relate to objects, places, people, and activities, reflecting what Tulving termed as 'episodic memory'. Consequently, the flexible search and retrieval through an intuitive language interface increase the complexity of NLQ by introducing significant challenges: handling videos of arbitrary length and integrating visual and textual modalities effectively. By treating the video as a text passage and the target moment as the answer span, the NLQ task shares significant similarities with span-based question answering (QA). However, while the causal relationships between word spans or sentences are often indirect and can be far apart, many events in a video are directly correlated and can even cause one another. Following the work of Zhang et al. [2], we model this task as a regression problem within a standard QA framework, utilizing video span localizing networks: VSLBase and its enhanced version, VSLNet. VSLNet addresses the aforementioned differences between video and text by incorporating a query-guided highlighting (QGH) strategy: it searches for the target moment within a highlighted region, thereby focusing on subtle differences between video frames.

To reduce the computational cost of training and increase performance, models for video moment localization are typically initialized with powerful and flexible fixed feature extractors for both textual and visual features. For the visual features, we conducted experiments using two pre-trained models with different domain adaptations. The first model, EgoVLP [3], is a video-language model specifically designed for video-language tasks on egocentric videos. It was trained on a generated dataset from Ego4D, which includes corresponding textual descriptions. The second model, Omnivore [4], was trained simultaneously on multimodal vision inputs, while not specifically on egocentric videos, demonstrating strong domain adaptation to general visual modalities. For the text features, we experimented with both static and dynamic word embeddings that are well-known in natural language processing (NLP): the former using the GloVe model [5] and the latter using the

Category	Template
Objects	Where is object X before / after event Y? Where is object X? What did I put in X? How many X's? (quantity question) What X did I Y? In what location did I see object X ? What X is Y? State of an object Where is my object X?
Place	Where did I put X?
People	Who did I interact with when I did activity X? Who did I talk to in location X? When did I interact with person with role X?

Figure 1. Query templates from the Ego4D NLQ benchmark [1]

BERT model [6].

To further enhance the practical applicability of the NLQ task, the second phase of our work focuses on generating text answers from predicted video segments based on natural language queries. By formulating this as a video-question answering task, it requires a model with robust comprehension abilities to follow human-provided instructions represented by the complex queries in the NLQ task. Additionally, the model must possess a strong capacity for video understanding according to these instructions. Recently, large language models (LLMs), such as GPT-3.5, GPT-4 [7], and Vicuna [8], have exhibited the comprehension ability to respond effectively to human instructions in input text. Large Vision Language Models (LVLMs) address the challenges of video language understanding by leveraging a powerful visual encoder and a projection layer to map visual input into text-like tokens as input for an LLM model. Therefore, this phase of the project involves the extraction of textual answers from identified video segments using an LVLM model called Video-LLaVA. [9].

2. Related Work

2.1. Egocentric Video Datasets

The release of datasets like EPIC-KITCHENS and Ego4D has significantly advanced research in egocentric vision. EPIC-KITCHENS [10] focuses on capturing daily kitchen activities, offering annotated videos for tasks such as action recognition and anticipation. In contrast, Ego4D [1], the dataset used in our project, is more extensive, encompassing a diverse array of daily activities across various environments. This broader scope greatly enhances the potential for egocentric video analysis. Both datasets provide rich annotations, including natural language queries, which are essential for developing and evaluating models that integrate visual and textual data.

2.2. Video Moment Localization

Video moment localization aims to identify a target segment within a video described by a given natural language query. In supervised learning, the first works, such as the one by Gao et al. [11], treat the video moment localization task as a matching problem, selecting the moments with best score among moment candidates. Additionally, another multimodal matching architecture employs a Temporal Adjacent Network (2D-TAN) [12], which selects the optimal moment from a temporal 2D map where moments are mapped according to their temporal distances. Although these models represent an early architecture, they require dense sampling of candidate moments to achieve good performance, and are sensitive to negative samples, which lead to low efficiency and a lack of flexibility. An architecture based on attention-like mechanism, coupled with the direct regression of the boundaries of the video segment, overcomes this drawbacks, as proposed by Zhang [2]. Specifically, this work illustrates the potential of two video span localizing networks based on a standard QA framework, which treat the video as text passage and the target moment as the answer span: VSLbase and its enhanced version, VSLNet. Moreover, the article adresses the limits of QA approach in video moment localization, adding a Query-Guided-Highlighting in VSLNet architecture to overcome the baseline model, VSLbase.

2.3. Large Vision Language Model

Large Language Models (LLMs), such as GPT-4 [7] and Meta Llama [13], rely on powerful comprehension abilities to follow human text instructions based on image or text inputs. Large Vision Language Models (LVLMs), leveraging the robust reasoning capabilities of LLMs, extend the modality interactions to videos. Most existing approaches to LVLMs, such as Macaw-LLM [14] and X-LLM [15], encode images and videos into separate feature spaces, which are then fed as inputs to LLMs. However, due to the lack of unified tokenization for images and videos, specifically due to misalignment before projection, as mentioned in the Video-LLaVA article [9], it becomes challenging to learn cross-modal interactions. Consequently, their performance on video understanding tasks falls significantly behind that of specialized video models, such as Video-ChatGPT [16]. The new baseline LVLM, Video-LLaVA [9], addresses this issue by aligning video and image data within a common feature space before projecting them into the LLM. As illustrated in the LLaVA paper [9], the Video-LLaVa model benefits from this unified representation, outperforming models specifically designed for video tasks, such as Video-ChatGPT.

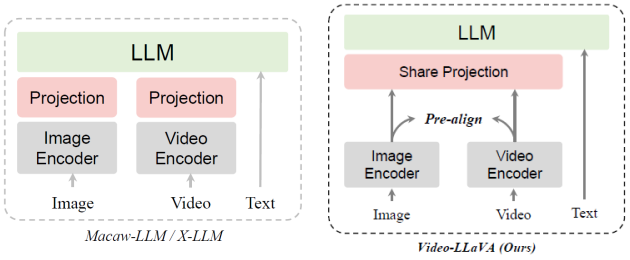

Figure 2. Video-LLaVA aligns images and videos before projection, allowing LLM to learn from a unified visual representation and endowing LLM with the ability to comprehend both images and videos simultaneously [9].

3. Methodology

3.1. Natural Language Video Localization

3.1.1 Problem Formulation

Given a language query $Q = \{q_j\}_{j=1}^{m}$ and an untrimmed video $V = \{f_j\}_{t=1}^{T}$, where m and T are the number of words and frames, respectively, the NLQ task requires the computation of a^s and a^e, which represent the start and end times of the relevant video segment where the query is answerable.

3.1.2 VSLBase

VSLBase is composed of two branches that extract visual and textual features from the input video V and the corresponding textual query Q, as represented in Figure 3. Fixed pre-trained models extract the features from the text and visual input Q and V, avoiding an unfeasible end-to-end training of the model. For each video V in the Ego4D dataset, we have pre-extracted visual features $\mathbf{V}$, while for each text query Q, the model employs a word embedding like GloVe or BERT to extract $\mathbf{Q}$ as a set of token vectors. After the projection of feature vectors in the same dimension, a shared encoder block maps the features of both modalities in a shared features space. This encoder block consists of four convolutional layers followed by a multi-head attention layer [17]. After feature encoding, the Context-Query Attention module (CQA) computes the similarity between visual and textual features using an attention-like mechanism. Finally, the conditioned span predictor based on two unidirectional LSTMs regresses the probability distributions of start and end boundaries.

The training objective is defined as:

$$\mathcal{L}_{\text{span}} = \frac{1}{2}\left[f_{\text{CE}}(P_s, Y_s) + f_{\text{CE}}(P_e, Y_e)\right].$$

where $P_s \in \mathbb{R}^n$ and $P_e \in \mathbb{R}^n$ are the regressed probability distributions of start (a^s) and end (a^e) boundaries, respectively; f_{CE} represents the cross-entropy loss function; Y_s and Y_e are the labels for the start and end boundaries.

3.1.3 VSLNet

VSLNet extends VSLBase by adding a Query-Guided Highlighting (QGH) module, as shown in Figure 3, which classifies the target moment as the foreground and the rest as the background. Specifically, the QGH module, employing an attention-like mechanism, aligns the target moment with the language query. Subsequently, QGH extends the boundaries of the foreground to cover its antecedent and consequent video contents with an extension ratio controlled by a hyperparameter, as shown in Figure 4. The extended boundary could potentially cover additional contexts and also help the network to focus on subtle differences between video frames. The loss function of the QGH module is:

$$\mathcal{L}_{QGH} = f_{\text{CE}}(S_h, Y_h)$$

where S_h represents the highlighting score of the attention mechanism on the alignments between text and visual features, and Y_h is a vector of 0-1 obtained by assigning 1 to the foreground and 0 to the background.

Therefore, VSLNet is trained by minimizing the combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{span}} + \mathcal{L}_{QGH}.$$

3.2. Generation of a Textual Answer

Given a natural language query as textual input $\mathbf{X}_T$, and the predicted video segment (from NLQ task) as raw signal $\mathbf{X}_V$, the second phase of the project requires the generation of an textual answer to the query.

3.2.1 Video-LLaVA

As shown in Figure 5, Video-LLaVA employs LanguageBind encoders $f_{\mathbf{V}}$ [18] to extract features from the raw visual signals (e.g. images or videos), a large language model $f_{\mathbf{L}}$ such as Vicuna, a visual projection layer $f_{\mathbf{P}}$ and a word embedding layer $f_{\mathbf{P}}$. The LanguageBind encoders align images and videos with the textual feature space, allowing to learn a unified visual representation within the model. The unified visual representation is fed into LLM after passing through a shared projection layer.Thus, the LLM is cabable of learning a visual understanding from a unified representation. The process of generating responses is comparable to that of a LLM. The input signals are encoded as a sequence of tokens:

$$\mathbf{Z_T} = f_{\mathbf{T}}(\mathbf{X_T}), \quad \mathbf{Z_V} = f_{\mathbf{P}}(f_{\mathbf{V}}(\mathbf{X_V})).$$

The model achieves multimodal understanding capabilities and generates a response sequence $f_{\mathbf{X_A}}$ of lenght L by max-

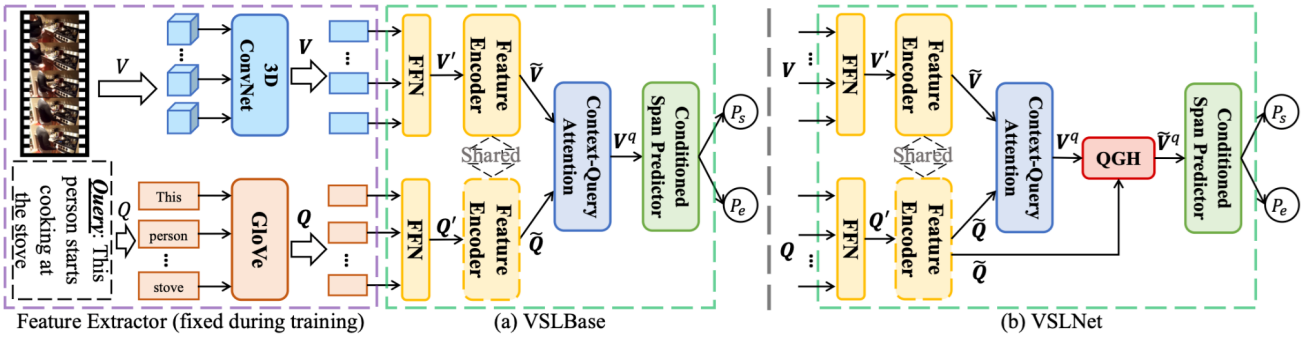

Figure 3. Architecture of the VSLBase and VSLNet models [2]

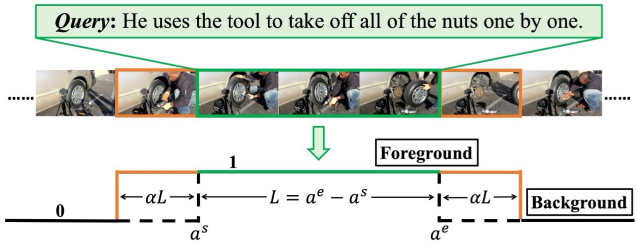

Figure 4. An illustration of the foreground and background of visual features. α is the ratio of foreground extension [2].

imizing the likelihood probability as in the following equation:

$$p(\mathbf{X_A}|\mathbf{X_V}, \mathbf{X_T}) = \prod_{i=1}^{L} p_\theta \left(\mathbf{X_A^{[i]}}|\mathbf{Z_V}, \mathbf{Z_T^{[1:i-1]}}\right),$$

where θ is a trainable parameter. Moreover, Video-LLava is characterized by a joint training on images and videos. The training pipeline can be divided into two parts: in the fist one, the model learns the video understanding, while in the second involving LLM in instruction tuning learn the spanning of answer in the conversation for video understanding task.

4. Experiments

4.1. Dataset

We conduct experiments on the NLQ benchmark of the Ego4D dataset. The NLQ annotations are from 227 hours of video, which cover 34 spatial scenarios, with a total of $19.2K$ queries spanning the selected 13 queries templates. The dataset's diverse and extensive nature provides a robust foundation for training and validating our models. The queries follow specific templates, as shown in Figure 1, and may relate to objects, places, people, and activities, reflecting the past experience of head-camera users.

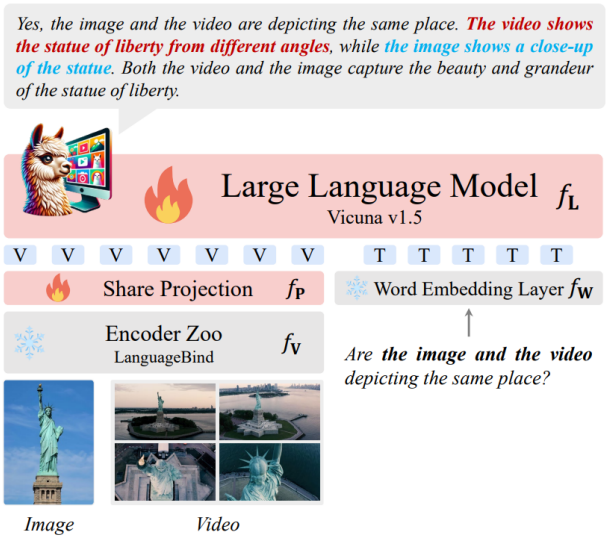

Figure 5. Illustration of Video-LLaVA [9]

4.2. Experimental Setup

We trained and validated VSLBase and VSLNet on the NLQ benchmark dataset of Ego4D, initializing the network with different fixed feature extractors. For visual features, we used pre-extracted features from pre-trained models EgoVLP and Omnivore. For text features, we employed BERT and GloVe word embedding models for extraction. All training processes were performed with a batch size value of **32**, a number of epochs of **10** and an initial learning rate of **0.0025** in an Adam optimizer. Performance on the Ego4D NLQ benchmark is evaluated using top-k recall at a specific temporal Intersection over Union (tIoU) threshold. Specifically, we report the results for $k \in \{1, 5\}$ and $m \in \{0.3, 0.5\}$, where m represents the threshold value. This metric is defined as the percentage of language queries that have at least one correct moment retrieval among the top-k retrieved moments.

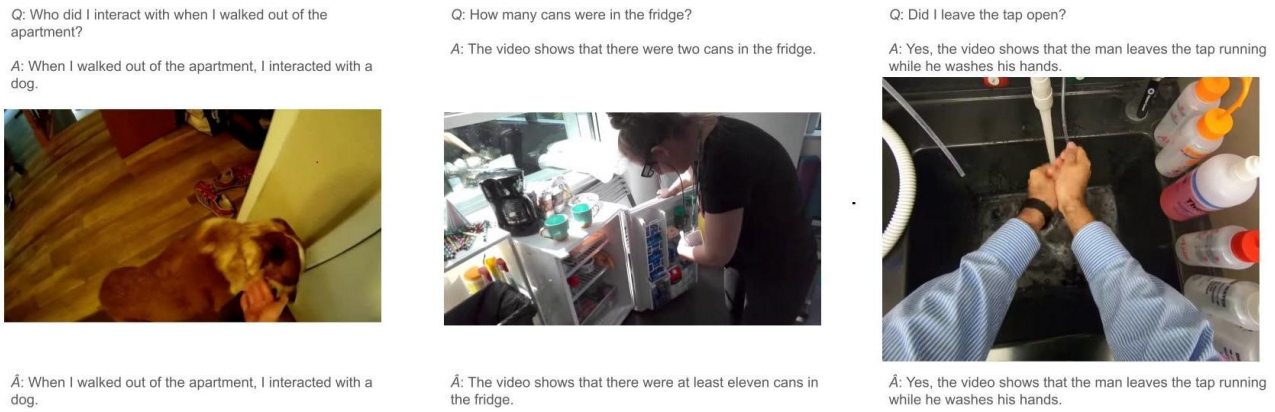

Figure 6. Video-LLaVA performance examples. A is the Video-LLaVA answer; $\hat{A}$ is the ground truth answer.

In the second phase of the project, we selected 50 NLQ queries and their related video segments retrieved by the best-performing model from the previous step: the VSLNet model configured with EgoVLP-BERT. The queries were selected based on the model's descending tIoU scores of the video segments. In this phase, we employed Video-LLaVA on the selected video segments and their related NLQ queries to generate textual answers. The answer predictions are evaluated with the Blue score [19] with respect to the ground truth answer.

4.3. Experimental Results

Results are shown in Table 2. As expected, all models performs better on EgoVLP features than on Omnivore features across all evaluations since EgoVLP is specifically designed for egocentric video-language understanding, leveraging a pretraining on a sample of Ego4D. While Omnivore, is a visual multimodal model , not specific trained on egocentric video and video language understanding task. For text features, BERT outperforms GloVe across all evaluations. BERT's transformer-based architecture enables it to understand the context of words within sentences more effectively than the static word embeddings provided by GloVe. Our best-performing trained model outperforms the baseline model of the NLQ task.

For Video-LLaVA answer generations, we obtained an average value of **0.488747094** for their BLEU scores. Thus, Video-LLaVA demonstrated a mix of accurate and inconsistent answers. It correctly identified and described the interaction between a person and objects or animals, and specific actions performed (as shown in the third frame in Figure 6). However, it presents problems in object recognition and answering quantitative queries (as shown in the central frame in Figure 6).

5. Conclusion

In this work, we investigated the use of natural language queries (NLQ) for egocentric video understanding using the extensive Ego4D dataset. The project consisted of two phases: video moment localization and generating textual answers based on identified moments. In the first phase, VSLNet, with the feature configuration EgoVLP-BERT, achieved the best performance in our experiments, outperforming the baseline model in the NLQ benchmark and demonstrating that the Query-Guided Highlighting (QGH) module in VSLNet significantly improved the ability to identify relevant video segments. This result illustrates the significant role of the QGH module in VSLNet, enhancing its ability to identify relevant video segments.

In the second phase, we used the Video-LLaVA model to generate textual answers from the identified video segments. While this approach showed potential, the performance indicated that there is significant room for improvement in generating contextually accurate answers.

Overall, this project demonstrates significant progress toward developing an augmented reality assistant capable of interpreting daily-life egocentric videos and responding to natural language queries.

References

[1] K. Grauman, A. Westbury, E. Byrne, *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.

[2] H. Zhang, A. Sun, W. Jing, *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.

[3] K. Q. Lin, A. J. Wang, M. Soldan, *et al.*, "Egocentric video-language pretraining," *Advances in Neural*

Methods	**Video-text Pre-extracted Features**		**IoU**=0.3		**IoU**=0.5	
	Video	Text	r@1	r@5	r@1	r@5
VSLBase	Omnivore	BERT	5.19	11.62	2.94	7.41
VSLBase	EgoVLP	BERT	6.14	12.47	4.03	8.42
VSLNet	Omnivore	BERT	6.48	14.04	3.72	8.70
VSLNet	Omnivore	GloVe	3.59	9.16	1.88	5.16
VSLNet	EgoVLP	BERT	**7.46**	**15.64**	**4.59**	**10.64**
VSLNet	EgoVLP	GloVe	3.67	10.02	2.07	6.07

Table 1. Recall for several IoUs on the NLQ task

Method	**Baseline**	**IoU**=0.3		**IoU**=0.5	
		r@1	r@5	r@1	r@5
VSLNet	Ego4D	5.45	10.74	3.12	6.63

Table 2. Recall for several IoUs on the NLQ task in the Ego4D baseline

Information Processing Systems, vol. 35, pp. 7575–7586, 2022.

[4] R. Girdhar, M. Singh, N. Ravi, *et al.*, "Omnivore: A single model for many visual modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 102–16 112.

[5] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.

[6] J. Devlin, M.-W. Chang, K. Lee, *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[7] O. J. Achiam, S. Adler, S. Agarwal, *et al.*, "Gpt-4 technical report," 2023.

[8] W.-L. Chiang, Z. Li, Z. Lin, *et al.*, *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality*, 2023.

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

[10] D. Damen, H. Doughty, G. M. Farinella, *et al.*, "Rescaling egocentric vision," *International Journal of Computer Vision*, pp. 1–23, 2022.

[11] J. Gao, C. Sun, Z. Yang, *et al.*, "Tall: Temporal activity localization via language query," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5277–5285.

[12] S. Zhang, H. Peng, J. Fu, *et al.*, "Learning 2D temporal adjacent networks for moment localization with natural language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020.

[13] H. Touvron, T. Lavril, G. Izacard, *et al.*, *Llama: Open and efficient foundation language models*, 2023.

[14] C. Lyu, M. Wu, L. Wang, *et al.*, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," *ArXiv*, vol. abs/2306.09093, 2023.

[15] F. Chen, M. Han, H. Zhao, *et al.*, "X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages," *arXiv preprint arXiv:2305.04160*, 2023.

[16] M. Maaz, H. A. Rasheed, S. H. Khan, *et al.*, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *ArXiv*, vol. abs/2306.05424, 2023.

[17] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[18] B. Zhu, B. Lin, M. Ning, *et al.*, "Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment," *ArXiv*, vol. abs/2310.01852, 2023.

[19] K. Papineni, S. Roukos, T. Ward, *et al.*, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318.