Lyric-Aware Karaoke Background Video Selection Using Large Language Models and Moment Retrieval

Tomoki Ariga*, Jun Taniguchi*, Yosuke Higuchi*, Sayaka Toma*, Kunihiro Abe†, Rie Shigyo†, and Tetsuji Ogawa*
* Department of Communications and Computer Engineering, Waseda University, Tokyo, JAPAN
E-mail: jtaniguchi@pcl.cs.waseda.ac.jp
† DAIICHIKOSHO CO.,LTD., Tokyo, JAPAN

***Abstract*—We propose a method for constructing coherent background videos by automatically selecting and combining video segments that semantically align with a song's lyrics. Rather than simply visualizing the literal content of the lyrics, the selected segments aim to convey scenes and atmospheres inferred through a deeper interpretation of their meaning. In the proposed approach, a large language model is first used to interpret the lyrics in terms of multiple attributes, such as characters, season, time of day, weather, location, and mood. These interpretations then serve as queries for moment retrieval, enabling the extraction of relevant video segments. To further optimize the composition of segments assigned to each part of the song (e.g., verse, bridge, chorus), a re-ranking step based on a second round of moment retrieval is applied to the initial retrieval results, ensuring global consistency across the final background video. The effectiveness of the proposed automatic video selection method based on lyric interpretation is validated through a subjective evaluation experiment comparing the generated videos with those used in actual karaoke systems.**

I. Introduction

In karaoke background video production, it is common practice to select video clips that match the lyrics and overall mood of the song from a vast pool of footage and assign them to each part of the song. However, this production process requires specialized expertise, entails significant time and financial costs, and may lead to biases in the selection of video content. Consequently, there is a strong demand for technologies that can automatically select suitable video segments and efficiently generate diverse background videos.

In this study, we propose a method for automatically extracting video segments that align with the content of song lyrics from a given set of candidate background videos. Moment retrieval models are widely employed to localize segments (moments) within videos that correspond to given text queries [1]–[4], and recent advances have focused on technical improvements as well as the development of related tasks that extend their applicability [5]–[8]. However, conventional moment retrieval models are primarily designed to localize video segments where the subject actions described in the query are directly depicted.

In contrast, karaoke background videos are typically expected not to portray the literal meaning of the lyrics, but rather to convey scenes and atmospheres inferred through a semantic interpretation of the lyrics. Directly applying existing moment retrieval models to such interpretive queries is therefore insufficient. Furthermore, methods for integrating retrieved video segments that correspond to different parts of a song, such as verses, bridges, and choruses, into a single coherent background video have not been thoroughly explored.

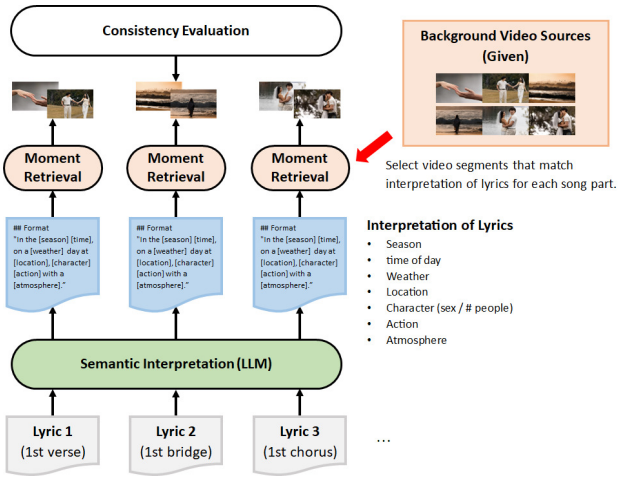

Fig. 1: Overview of video segment selection model. Model consists of three components: 1) lyric interpretation module that performs semantic analysis of lyrics, 2) moment retrieval module that selects relevant video segments based on interpreted content, and 3) coherence evaluation module that determines optimal combination of segments by assessing overall consistency of resulting video.

To address these challenges, this study develops a moment retrieval framework in which song lyrics are semantically interpreted by a large language model (LLM) [9], [10] to generate attribute information relevant to background video selection, such as characters, season, time of day, weather, location, and mood, and these interpretations are then used as retrieval queries. Additionally, we explore a re-ranking strategy to optimize the combination of video segments assigned to each part of the song, thereby enhancing the overall consis-

979-8-3315-7206-8/25/$31.00

tency of the final background video.

The effectiveness of the proposed method is validated through a subjective evaluation comparing the automatically generated background videos with those used in commercial karaoke systems. The findings of this research contribute to a novel video retrieval framework that focuses not on literal text matching but on deeper semantic interpretation of lyrical content.

The remainder of this paper is organized as follows. Section II provides a detailed description of the proposed method for selecting background videos that align with song lyrics. Section III presents the subjective evaluation experiment conducted to assess the effectiveness of the proposed approach and discusses the findings. Finally, Section IV concludes the paper with a summary of this study.

II. Generation of Background Videos Aligned with Song Lyrics

In this study, we focus on the scenario in which the main video sources have already been selected from an extensive video library, and appropriate scenes are then assembled according to the lyrics to create the final background video. We investigate an automatic method for determining which segments of these sources should be assigned to each part of the song.

Figure 1 illustrates the overall architecture of the proposed automatic background video selection framework. The system comprises three main modules:

1) a *lyric interpretation module* that generates interpretation sentences containing attribute information from the lyrics using an LLM;
2) a *moment retrieval module* that extracts relevant scenes from background video sources based on the generated interpretations; and
3) a *consistency evaluation module* that assesses the coherence among the retrieved video segments.

By integrating these modules, the framework enables the automatic generation of a single background video that aligns with the lyrical content of the song.

The remainder of this section provides detailed explanations of the lyric interpretation module (Section II-A), the moment retrieval module (Section II-B), and the consistency evaluation module (Section II-C).

A. Lyric Interpretation Using Large Language Model

When selecting background video segments to accompany song lyrics, it is generally preferable to avoid literal visualizations. Instead, it is desirable to select scenes and atmospheres that reflect an interpretation of the lyrics' meaning. Based on this perspective, our approach interprets song lyrics as a narrative described by seven attributes: *season*, *time of day*, *weather*, *location*, *characters*, *actions*, and *mood*.

Among these attributes, six, excluding *actions*, are used to annotate the background video sources, consistent with conventional karaoke video production practices. Accordingly, these six attributes are expected to serve as effective cues for aligning the semantic content of the lyrics with the available video material.

```
You will be given the lyrics and title of a song.
Please analyze the lyrics of each part according to the
    following format.
Be sure to include the season, time of day, weather,
    character, and their actions, as well as the
    atmosphere, in a concise sentence.
Ensure that each part connects to form a coherent overall
    story.
Do not simply describe the literal content of the lyrics;
    Instead, describe the actions and scenes that are
    implied or associated with them.
Avoid emotional or abstract expressions, and focus on
    concrete actions.

## Format for Analysis
Part: 1st verse, a-melody
Result: "In the [season] [time], on a [weather] day at [
    location], [character] [action] with a [atmosphere]."
Part: 1st verse, b-melody
Result: ...

## Example Format of Given Lyrics
Title: (Song Title)
1st verse, a-melody:
(Lyrics of the first verse, a-melody)
1st verse, b-melody:
(Lyrics of the first verse, b-melody)
1st verse, chorus:
(Lyrics of the first verse, chorus)

## Example of Analysis
Part: 1st verse, a-melody
Result: In the winter morning, on a snowy day at the park,
     men and women meet and then part ways with a
    negative atmosphere.
```

Fig. 2: Prompt for lyric interpretation.

To generate interpretation sentences that incorporate these attribute details, we employ an LLM, specifically GPT-4o [9]. Figure 2 shows an example of the prompt used for this purpose. This prompt consists of three parts: an instruction, an output format definition, and illustrative examples.

The **instruction** directs the LLM to interpret the lyrics as a concrete narrative covering all seven attributes, with a particular emphasis on the *actions* to maintain consistency across selected video segments. The **output format definition** specifies that the song title and all lyric segments are given as input, and that the model should output an interpretation sentence for each segment. The **illustrative examples** provide reference analyses that help guide the LLM to produce interpretation sentences consistent with the desired structure.

B. Video Segment Selection Using Moment Retrieval

In the proposed framework, moment retrieval is performed for each section of a song (e.g., verse, bridge, chorus) to identify the most appropriate video segments, which are then concatenated to generate a coherent background video. This section provides an overview of the moment retrieval technique and explains how it is adapted for lyric-based queries.

1) Moment Retrieval: Moment retrieval is a technique that locates segments in a video that correspond to a given text query. While early methods typically relied on separate scoring mechanisms, recent advances have enabled unified, end-to-end

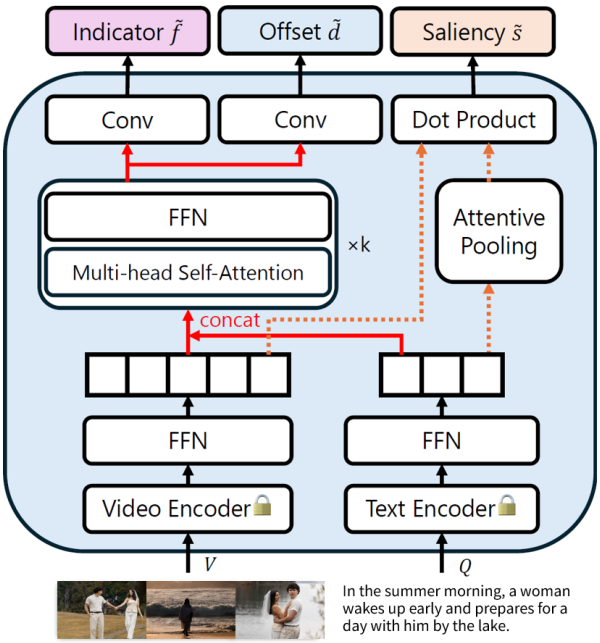

Fig. 3: Architecture of moment retrieval model.

approaches [3], [4]. In this study, we adopt a model based on UniVTG [4], which provides a unified framework capable of handling various Video Temporal Grounding (VTG) tasks.

As illustrated in Fig. 3, UniVTG produces three types of predictions via separate output heads: foreground indicator $f_i \in \{0, 1\}$, boundary offsets $d_i = [d_i^s, d_i^e] \in R^2$, and saliency score $s_i \in [0, 1]$.

Foreground head (for matching) $\tilde{f}$: Each clip v_i is classified as either foreground (relevant to the text query Q) or background (irrelevant). The loss for this head is defined using Binary Cross Entropy (BCE) between the predicted label $\tilde{f}_i$ and the ground-truth label f_i, as follows:

$$\mathcal{L}_f = -\lambda_f (f_i \log \tilde{f}_i + (1 - f_i) \log(1 - \tilde{f}_i)), \tag{1}$$

where λ_f denotes the weighting coefficient for the foreground loss.

Boundary head (for localization) $\tilde{d}$: This head predicts the start and end timestamps of the relevant video segment. Let t_i be the center timestamp of each clip v_i, and $b_i = [t_i - d_i^s, t_i + d_i^e]$ denote the start and end of that segment. The loss is defined as the sum of the L1 prediction error $\mathcal{L}_{\text{L1}}$ between the predicted offsets $\tilde{d}_i$ and the ground truth d_i, and the overlap error $\mathcal{L}_{\text{IoU}}$ between the predicted boundary $\tilde{b}_i$ and the ground truth b_i:

$$\mathcal{L}_b = 1_{f_i=1}[-\lambda_{\text{L1}} \mathcal{L}_{\text{L1}}(\tilde{d}_i, d_i) + \lambda_{\text{IoU}} \mathcal{L}_{\text{IoU}}(\tilde{b}_i, b_i)], \tag{2}$$

where $1_{f_i=1}[\cdot]$ is an indicator that activates this loss only for clips labeled as foreground. λ_{L1} and λ_{IoU} are the weighting coefficients for each term.

Saliency head (for contrasting) $\tilde{s}$: This head predicts the saliency score of each clip, indicating its relevance to the text query. The score is computed using the cosine similarity between the clip feature $\mathbf{v}_i$ and the query embedding $\mathbf{S}$:

$$\tilde{s}_i = \frac{\mathbf{v}_i^{\text{T}} \mathbf{S}}{|\mathbf{v}_i|_2 |\mathbf{S}|_2}. \tag{3}$$

To learn effective clip-to-query matching, contrastive learning is applied both within videos (intra-video) and across different videos (inter-video). For intra-video contrastive learning, a foreground clip v_p is randomly sampled, and clips v_j with lower saliency scores $s_j < s_p$ are treated as negatives i.e., $\Omega = \{j | s_j < s_p, 1 \leq j \leq L_v\}$, where L_v denotes the number of video clips. The intra-video loss is defined as:

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

where τ denotes a temperature parameter.

For inter-video contrastive learning, queries from other samples in the batch act as negative samples, with the loss defined as:

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

where B denotes the training batch size and $\tilde{s}_p^k = \cos(\mathbf{v}_i, \mathbf{S}_k)$.

The final saliency loss combines the intra- and inter-video losses:

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

The overall objective function sums the foreground, boundary, and saliency losses:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (\mathcal{L}_f + \mathcal{L}_b + \mathcal{L}_s), \tag{7}$$

where N is the number of clips in the training batch.

For implementation, we use CLIP [11] (ViT-B/32) as both the video and text encoders. At inference time, predicted boundaries $\{\tilde{b}_i\}_{i=1}^{L_v}$ are ranked according to their saliency scores $\tilde{s}_i$. Since predicted segments may overlap, Non-Maximum Suppression (NMS) is applied to remove redundant intervals before final ranking.

2) Retraining of Moment Retrieval Model for Lyric Interpretation Descriptions: For the task of selecting video segments that appropriately match song lyrics, it is essential to consider not only the actions depicted in the video but also a range of contextual attributes, such as season, time of day, weather, location, and mood. However, conventional moment retrieval methods primarily focus on subject actions, and queries in standard training datasets for moment retrieval rarely include rich contextual information like season or weather.

To address this limitation, we construct a custom dataset to enable moment retrieval using queries derived from lyric interpretations and use it to retrain the moment retrieval model. Specifically, we automatically extract six key attributes, character, season, time of day, weather, location, and mood, from videos in the QVHighlights [3] dataset, which is widely used for training moment retrieval systems. This attribute extraction

is performed using LLaVA-NeXT [12], [13], a state-of-the-art vision-language model (VLM).

We then combine this automatically extracted attribute information with the action annotations originally provided in QVHighlights to generate unified attribute descriptions using an LLM. By retraining the moment retrieval model with these enriched descriptions, the system can retrieve video segments based on queries that more accurately reflect the nuanced interpretations of song lyrics.

The prompts used for annotation follow the same three-part structure as those used for lyric interpretation: an instruction, a clear output format specification, and illustrative examples. Additionally, when an attribute cannot be determined from the video, the system is explicitly instructed to output "None" to avoid assigning ambiguous or spurious labels.

C. Composition of Retrieved Video Segments via Re-ranking

Background video segments retrieved based on lyric interpretations are first extracted for each part of the song, such as the verse, bridge, and chorus. These part-wise segments must then be integrated into a single, coherent background video. However, simply concatenating the top-ranked segments for each part does not necessarily yield an optimal or globally consistent result.

To address this, we introduce a second round of moment retrieval on candidate background videos formed by concatenating the initially retrieved segments. If the results of this second retrieval align well with those of the first, the composition is regarded as maintaining overall coherence. The final background video is then determined by selecting the combination of segments that best preserves this consistency.

The procedure consists of the following steps:

1) **Extraction of candidate segments**: Perform moment retrieval and extract the top N candidate segments for each song part.
2) **Generation of candidate videos**: Construct all possible combinations of the extracted segments across the song parts to generate candidate background videos.
3) **Final selection**: Perform moment retrieval on each candidate background video using the original text queries and select the candidate with the highest final score as the final output.

The final score for each candidate video is computed based on two factors: the top-ranked saliency score S^p_{saliency} obtained from moment retrieval for the text query Q_p of each song part p, and the overlap ratio S^p_{overlap} between the segments retrieved before and after re-ranking. The final score is calculated as:

$$S = \sum_p S^p_{\text{overlap}} \cdot S^p_{\text{saliency}}. \tag{8}$$

III. Background Video Selection Experiment

To evaluate the effectiveness of the proposed method, we conducted a subjective assessment experiment. Specifically, we compared the background videos actually used in karaoke systems with those generated by our method.

TABLE I: List of questionnaire items

Questions	Contents
Q1	No sense of discomfort while watching the video
Q2	The mood of the video matches that of the song
Q3	The theme of the video matches the lyrics
Q4	The storyline of the video aligns with the lyrics
Q5	The time of day, weather, and season in the video match the lyrics
Q6	The transitions between video segments feel natural
Q7	Comments on any parts that felt particularly mismatched or uncomfortable, and the reasons

TABLE II: Cronbach's coefficient α.

	GT	Top-1	Re-ranked
song 1	0.755	0.885	0.846
song 2	0.853	0.717	0.738
song 3	0.866	0.471	0.833
song 4	0.806	0.860	0.728
song 5	0.806	0.860	0.728
average	0.817	0.759	0.775

A. Experimental Setups

A total of 14 participants took part in the experiment. Each participant evaluated how well the background videos matched the lyrics for the same set of five songs. For each song, participants compared three types of background videos:

1) **Ground-truth**: The actual video used in karaoke systems.[1]
2) **Top-1**: A video composed by concatenating the top-ranked segments for each song part based on moment retrieval.
3) **Re-ranked**: A video composed by concatenating segments selected using the re-ranking scores.

The underlying moment retrieval model was based on UniVTG. It was pre-trained on multiple VTG task datasets, including QVHL [3], Charades [14], NLQ [15], TACoS [16], ActivityNet [17], and DiDeMo [18], and then fine-tuned on an augmented version of QVHL containing text queries derived from lyric interpretations.

Participants rated each video using the seven items shown in Table I. Items Q1–Q6 were rated on a five-point Likert scale [19], while Q7 was open-ended. Lyrics were displayed as subtitles during the evaluation, and the order of video presentation was counterbalanced to mitigate order bias. The questions covered the overall impression (Q1–Q3) and the degree of consistency between the lyrics and the video content (Q4–Q6).

B. Experimental Results

[1]The Ground-truth videos are not exactly identical to the versions used in commercial karaoke systems. While the selected video segments from the source material are the same, the actual karaoke videos often include additional visual effects at segment transitions or insert short clips to enhance visual appeal. This study focuses exclusively on the segment selection process itself, and the evaluation was conducted using videos without such post-processing.

TABLE III: t-test results for subjective evaluation scores (p-values).

	Q1	Q2	Q3	Q4	Q5	Q6
GT vs. Top-1	0.115	0.142	0.348	0.329	0.042	0.045
GT vs. Re-ranked	0.371	0.249	0.420	0.053	0.058	0.172
Top-1 vs. Re-ranked	0.191	0.020	0.225	0.012	0.50	0.169

1) Questionnaire Reliability: The reliability of the subjective evaluations was assessed using Cronbach's coefficient α [20]. Table II shows the coefficient values for each video. For the **Ground-truth** videos, the coefficient exceeded 0.8, indicating high internal consistency. For the automatically generated videos (**Top-1** and **Re-ranked**), the coefficient was slightly lower but within an acceptable range (within 0.05 of 0.8), suggesting that the subjective responses remained reliable despite individual differences in preference.

2) Subjective Evaluation Results: Table III summarizes the t-test results for the subjective evaluation scores across the three video types.

For Q1–Q3, which assessed overall impression and naturalness, there was no significant difference between the automatically generated videos and the **Ground-truth** videos, indicating that the proposed method can produce videos with a natural look and feel. However, some participants pointed out that certain transitions still felt unnatural, indicating an area for further improvement.

For Q4, which evaluated storyline consistency, the **Re-ranked** videos achieved the highest scores, demonstrating the effectiveness of optimizing segment combinations through re-ranking.

For Q5, which examined whether the time of day and season matched the lyrics, the automatically selected videos scored significantly lower than the **Ground-truth** videos. This suggests that clearly conveying seasonal and temporal context remains a challenge, highlighting the need to improve the prompts provided to the LLM and to further refine the moment retrieval model.

For Q6, which evaluated the smoothness of segment transitions, the **Top-1** videos scored significantly lower than the **Ground-truth** videos, while the **Re-ranked** videos did not show a significant difference. This indicates that re-ranking contributes to smoother and more coherent transitions, although some awkward cuts still remain and should be addressed in future work.

IV. Conclusion

In this study, we proposed an automatic method for selecting karaoke background videos by combining lyric interpretation using an LLM with moment retrieval based on the interpreted descriptions. The effectiveness of the proposed approach was verified through a subjective evaluation experiment. The results showed that this method can generate background videos with a level of naturalness comparable to those used in karaoke systems by selecting video segments that closely align with the semantic content of the lyrics. Notably, the proposed approach received higher ratings than conventional karaoke videos for how well the visuals matched the song's narrative. Moreover, by optimizing the combination of video segments through re-ranking using a second round of moment retrieval, the method further improved visual continuity and storyline coherence across segment.

References

[1] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1247–1257.

[2] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 12 870–12 877.

[3] J. Lei, T. L. Berg, and M. Bansal, "QVHIGHLIGHTS: Detecting moments and highlights in videos via natural language queries," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 846–11 858, 2021.

[4] K. Q. Lin, P. Zhang, J. Chen, *et al.*, "UniVTG: Towards unified video-language temporal grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2794–2804.

[5] P. Li, C.-W. Xie, H. Xie, *et al.*, "Momentdiff: Generative video moment retrieval from random to real," *Advances in neural information processing systems*, vol. 36, pp. 65 948–65 966, 2023.

[6] X. Sun, J. Gao, Y. Zhu, X. Wang, and X. Zhou, "Video moment retrieval via comprehensive relation-aware network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 5281–5295, 2023.

[7] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-dependent video representation for moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 23 023–23 033.

[8] A. Zala, J. Cho, S. Kottur, *et al.*, "Hierarchical video-moment retrieval and step-captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 056–23 065.

[9] A. Hurst, A. Lerer, A. P. Goucher, *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[10] S. An, Z. Ma, Z. Lin, N. Zheng, J.-G. Lou, and W. Chen, "Make your LLM fully utilize the context," *Advances in Neural Information Processing Systems*, vol. 37, pp. 62 160–62 188, 2024.

[11] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021, pp. 8748–8763.

[12] H. Liu, C. Li, Y. Li, *et al.*, *LLaVA-NeXT: Improved reasoning, OCR, and world knowledge*, Jan. 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[13] Y. Zhang, B. Li, H. Liu, *et al.*, *LLaVA-NeXT: A strong zero-shot video understanding model*, Apr. 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

[14] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-Ego: A large-scale dataset of paired third and first person videos," *arXiv preprint arXiv:1804.09626*, 2018.

[15] 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, pp. 18 995–19 012.

[16] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos.," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013.

[17] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[18] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5803–5812.

[19] R. Likert, "A technique for the measurement of attitudes.," *Archives of Psychology*, 1932.

[20] L. J. Cronbach, "Coefficient alpha and the internal structure of tests.," *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.