Multilingual Video Grounding: Cross-Language Temporal Localization Without English Supervision

Ruthvik Kanumuri
Texas A&M University
kruthvik007@tamu.edu

Ramana Heggadal Math
Texas A&M University
ramana_hm@tamu.edu

Shravan Conjeevaram
Texas A&M University
shravan10@tamu.edu

Jnana Preeti Parlapalli
Texas A&M University
pj.preeti@tamu.edu

1 Introduction

Multilingual Video Grounding is the task of identifying temporal segments in a video that correspond to text queries from multiple languages. Most existing approaches that use video grounding rely on English as an intermediary crutch for representation learning, often involving translation or pretraining on English-language datasets (Sigurdsson et al., 2020). Although effective for well-resourced languages, this dependence becomes problematic for low-resource languages, where translations can be noisy or completely unavailable.

Sigurdsson et al. (2020) demonstrated that multilingual videos inherently contain shared semantic information, which can be leveraged to map words across languages using visual grounding. Their contrastive learning approach aligns text and video representations in an unsupervised manner. However, the translation length was limited to 4-5 words. We explored the possibility of using longer sentences (word length 9-10) while also avoiding the reliance on English.

Instead of generating full translations, we redefine the task of identifying and matching the closest semantically relevant phrases in the target language using visual context. This shift allows us to explore the effects of using longer sentences without burdening us with the complexities of word-by-word translation.

2 Related Work

Previous work on the Video Moment Retrieval has primarily focused on English datasets like TVR (Lei et al., 2020) and ActivityNet Captions (Krishna et al., 2017). Multilingual Video Grounding has been explored with datasets such as mTVR, which includes Chinese-English annotations (Lei et al., 2021).

Wang et al. (2024) have sought to improve cross-lingual video retrieval by using Multimodal Large Language Models (MLLMs) to generate visual descriptions, which are then used to align non-English text with video. However, these methods still incorporate English as an underlying bridge during training, limiting their applicability for true English-free supervision.

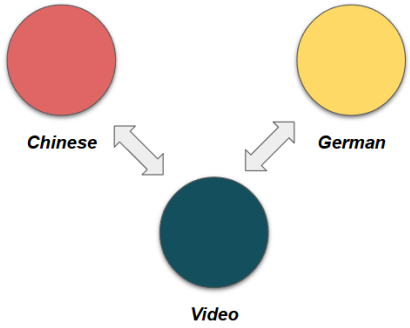

Figure 1: Learning Mappings for Embedding Spaces

Similarly, cross-lingual video-text models such as X-CLIP (Ma et al., 2022) introduce multilingual contrastive learning, but remain reliant on pretrained English models like CLIP (Radford et al., 2021). Existing literature lacks a widely established approach for directly mapping non-English text to video without relying on English. We aim to eliminate English as a pivot language, enabling direct alignment of non-English text with video representations and hoping to improve accessibility for truly low-resource languages.

3 Novelty and Challenges

3.1 Potential Key Innovations

- **Direct Multilingual Video-Language Alignment:** Our approach aims at learning direct text-video embeddings across semantically distant languages such as Chinese (Mandarin) and German. Unlike previous methods such as Sigurdsson et al. (2020), which used English as an intermediary for alignment, our

method avoids the dependency on English. This can improve accessibility to low-resource languages.

- **Supervised Contrastive Learning:** Within the bounds of phrase-matching, the model identifies natural cross-lingual correspondences and aligns them with video through our variation of contrastive learning in a shared semantic space. If the task were framed as word-by-word translation, a self-supervised approach would have been more appropriate.

- **Dual Encoder Framework for Multilingual Alignment:** To account for the nuances of different languages (Chinese and German), we follow a dual-encoder architecture, with separate Transformer models to encode each language. We understand that this architecture maps all inputs to a shared embedding space close to that of the video embeddings. Causally, an effective shared space can enable effective cross-lingual and cross-modal retrieval.

- **Triplet Loss for Improved Semantic Separation:** A variant of the contrastive loss, the Triplet loss aims to bring semantically similar data points closer, and push semantically dissimilar data points farther.

$$L_{triplet}(\alpha, p, n) = \max(0, \|\alpha - p\|_2 - \|\alpha - n\|_2 + m)$$

$$L_{reg}(a) = \|mean(a, dim = 0)\|_1$$

where α is the anchor (a reference sample among the video embeddings), p is the positive sample semantically similar to the anchor, and n is the negative sample semantically dissimilar from the anchor. Since we want to exacerbate the effect of larger differences in distance, we use the L_2 norm. In short, we try to bring α and p closer than α and n by at least a margin m. Regularization keeps the data points from collapsing onto each other when they are pulled together.

- **Pairwise Combination of Losses:** Since we are dealing with multiple languages and modalities at the same time, it is crucial to ensure that the loss is wholistic. This can be envisioned through bidirectional losses for all 2 mode permutations taken from the 3 original modes, namely Chinese (C), Video (V) and German (G). The total loss can be an average of the bidirectional losses.

$$L_{total} = \frac{1}{6}(L_{vc} + L_{vg} + L_{cg} + L_{cv} + L_{gv} + L_{gc})$$

Anchor	Positive	Negative
V	C	C → L_vc
V	G	G → L_vg
C	G	G → L_cg
C	V	V → L_cv
G	V	V → L_gv
G	C	C → L_gc

Figure 2: Pairwise Combination of Losses

3.2 Potential Challenges

- **Multimodality:** The alignment of heterogeneous data is complex because of its high dimensionality and varied nature of the data. In our work, this complexity is amplified by the inclusion of semantically distant languages (Chinese and German) and the absence of English-based supervision. Unlike static images, video content appears frame-by-frame over time, requiring the model to capture and condense temporal dependencies into fixed-length representations that somehow retain the essence. Furthermore, aligning three distinct modalities within a shared embedding space brings risks of semantic drift, especially under margin-sensitive loss functions like our triplet loss. These factors make our task highly non-trivial.

- **Translation granularity:** The traditional goal of sentence-level translation is replaced by retrieving the closest matching phrase in the target language. While this aligns better with self-supervised learning and avoids noisy translations, it introduces new challenges in retrieval granularity.

- **Phrase-Level Retrieval for Longer Sequences:** Although we focus on phrase retrieval, the model often handles longer spans (e.g., $\sim$10 words), increasing the difficulty of

achieving precise semantic alignment, especially when syntactic complexity varies between languages.

- **Semantic Divergence Across Languages:** Aligning languages with significant semantic and structural differences (e.g., Chinese and German) without using English as a bridge makes learning correspondences more complex and increases reliance on strong contrastive objectives.
- **Lack of Direct Non-English Datasets:** Most existing datasets like VATEX or mTVR rely on English anchors. The absence of high-quality, direct non-English-to-video data limits the development and evaluation of models aiming for English-free multilingual grounding.
- **Cross-Lingual Evaluation Metrics:** There is currently no standardized way to evaluate phrase-level alignment across languages. Defining fair and language-agnostic evaluation metrics remains a key challenge for benchmarking multilingual video grounding systems.

4 Approach

We have three main considerations in designing our approach. Namely, Dataset Formation, Loss Modeling, and Model Architecture

4.1 Dataset Formation

Languages
Chinese (Mandarin)
German

Table 1: Languages Considered

Dataset
VATEX (Chinese - English Video Captions)
VATEX (Machine Translated English - German)
VATEX Videos (sans captions)

Table 2: Generated Dataset

Given its availability of frame-level embeddings and multilingual captions, the *VATEX* dataset provided an ideal foundation for constructing our custom dataset. *VATEX* contains a large corpus of video frames with parallel Chinese and English captions. As it is, this setup will not suffice our needs, because we need parallel German captions. Since English is semantically similar to German, we machine-translated English text to German, and created our final corpus.

4.2 Loss Modeling

Figure 3: Transforming the Embedding Space

We explored two ways of mapping sentences from one language to another. Namely, learning mappings from one modality to another, and as an alternative, transforming the embedding space itself with contrastive learning. We found that the latter approach yielded better matches. Since handling 3 different modalities was challenging, we augmented the Triplet loss step with an additional step to garner pairwise mean of losses for all permutations of the 3 modalities, as mentioned earlier under Novelties and Challenges.

In the ideal scenario, our loss should cluster embedding groups on the basis of topic/genre/semantics and reduce the effect of modality on embedding separation. However, we anticipate practical limitations owing to which we settle for lesser overlaps and place more emphasis on proximity among the data points of the three spaces. To account for this practical limitation, we used distance as a metric to generate the final similarity score.

4.3 Architecting the Model

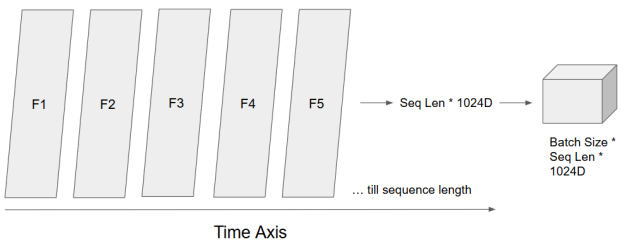

Figure 4: Input Representation - Video

Figure 4 showcases how we generate input representations from video frame embeddings. Each frame is represented by a 1024-D embedding. A collective representation of frames for the video

would be a stack of the frame embeddings, which in turn will be batched for processing.

Figure 5: Temporal Attention Block - Video

Figure 5 depicts video blocks being processed with Temporal Attention mechanism to garner inter-frame relationships while maintaining the same dimensions.

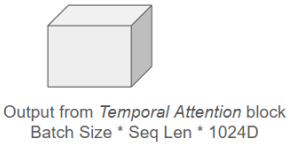

Figure 6: Mean Pooling

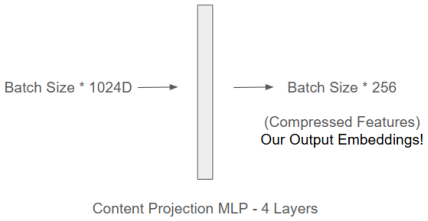

Figure 7: Compression

Figure 6 and 7 showcase our way of summarizing the frames to generate one representation per video, and compressing that summary from 1024D to 256D using a Multilayer Perceptron. Thus, we obtain dense embeddings for each video.

4.4 Text Processing and Optimization

Text captions are processed using language-specific encoders namely, XLM-Roberta (Figure 8) for slight English connotation, and BERT-base-Chinese and BERT-base-German (Figure 9) for completely avoiding English influence. The encoded embeddings were passed through the same type of content projection modules. Training was performed using mixed-precision with the Adam optimizer, cosine annealing learning rate scheduler, and early stopping. Both the contrastive and full triplet-supervised training variants were explored.

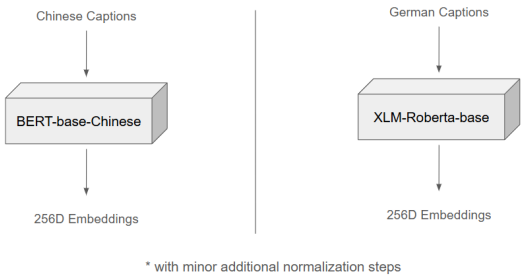

Figure 8: Text Processing(XLM-Roberta)

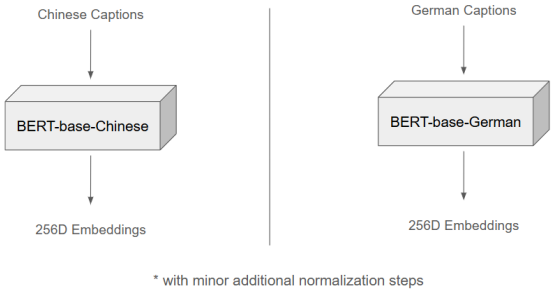

Figure 9: Text Processing(BERT-base-German)

5 Experimental Setting

1. **Inputs**: Video embeddings and tokenized Chinese and German captions.

2. **Video Embeddings Transformation**: Apply temporal pooling and an MLP to video features. Use BERT followed by projection layers for Chinese and German captions.

3. **Cross-modal Alignment**: Map Chinese and German embeddings to the video space, incorporating a cross-lingual transformation.

4. **Loss Computation**: Use pairwise triplet loss with regularization to align the modalities.

5. **Prediction**: Given a Chinese input, generate its embedding, find the nearest video embedding, and retrieve the top 3 matching German captions.

6. **Evaluation Protocol:** During inference, a Chinese caption is first embedded in the shared space. The model retrieves the nearest video embedding and subsequently identifies the top-3 closest German captions aligned with the retrieved video, simulating multilingual grounding and translation via video context.

6 Results, Findings, and Insights

- We chose **Recall@3** as our evaluation metric because our pipeline reduced the problem to phrase-matching using video summarization. We expected the correct match to appear among the top 3 results.
- Figures 10 and 11 show an example where a Chinese phrase about a hula hoop retrieves a correctly ranked German caption about a hula hoop.
- In contrast, for a Chinese phrase about a child running on a treadmill, the correct German caption appears among the top 3 but is ranked lower (Figures 12 and 13).
- This behavior was observed across both German encoders:
  - *XLM-Roberta*, which has slight English influence.
  - *BERT-base-German*, with no English influence.
- We hypothesize this is due to our video summarization approach. A single summary for a long video may not accurately capture all relevant contexts.
- Because our dataset spans many genres, our model struggled to generalize across all possibilities. A genre-specific model may have improved performance.
- Despite these challenges, our approach successfully used video as a cross-lingual bridge to connect semantically distant languages **without relying on English**.

Figure 10: Correct Predictions (Chinese-Expected German)

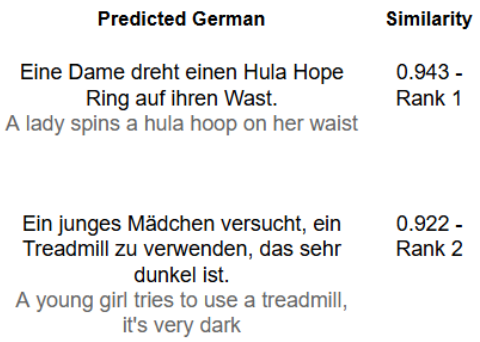

Figure 11: Correct Predictions (Predicted German-Similarity)

Figure 12: Misses (Chinese-Expected German)

7 Future Directions

- We believe that effective application of triplet loss requires more complex modeling, higher compute capacity, and larger datasets. A rerun of the experiment with amplified resources could yield better results.
- Instead of using a single summary vector per video, we could generate 2–3 summary vectors to better capture diverse contextual information.
- Conducting **genre-specific experiments** may help mitigate ranking setbacks by narrowing the domain and improving alignment accuracy.

References

R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J.C. Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715.

J. Lei, T.L. Berg, and M. Bansal. 2021. mvtr: Multilingual moment retrieval in videos. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 726–734.

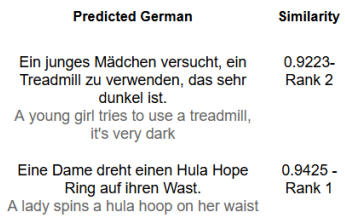

Figure 13: Misses (Predicted German-Similarity)

J. Lei, L. Yu, T.L. Berg, and M. Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Computer Vision Association (ECVA)*, pages 10814–10824.

C. Ma, X. Qian, J. Dong, C. Liu, X. Dai, Z. Liu, Y. Liu, Y. Li, X. Yu, F. Wu, and C. Chen. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. pages 1–18.

A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. 2021. Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning (ICML)*, 139:8748–8763.

G.A. Sigurdsson, J. Alayrac, A. Nematzadeh, L. Smaira, M. Malinowski, J. Carreira, P. Blunsom, and A. Zisserman. 2020. Visual grounding in video for unsupervised word translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13466–13476.

Y. Wang, L. Wang, Q. Zhou, Z. Wang, H. Li, G. Hua, and W. Tang. 2024. Multimodal llm-enhanced cross-lingual cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8296–8305.