

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

Ramana Heggadal Math

Ruthvik Kanumuri

Jnana Preeti Parlapalli

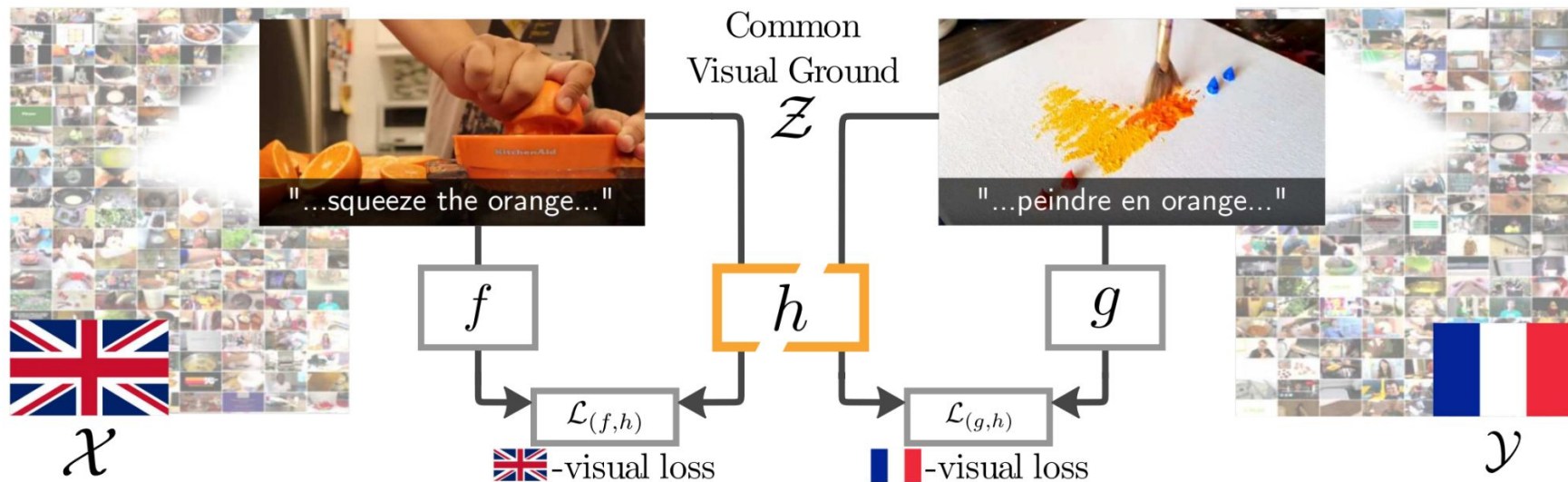
Shravan Conjeevaram



TEXAS A&M UNIVERSITY

Engineering

# Problem Statement



# ***Novelties and Challenges***

## **Novelties**

- Explicit mapping layers between Chinese-German, Text-Video. Base paper just had a visual-text space
- Loss Function: Triplet Loss for better separation and margin-based control
- Used Chinese-German as the base comparison, which is semantically very distant
- Dual encoders project into joint space
- Evaluation: Sentence level translation (matching)

## **Challenges**

- Original goal: To translate text by generating the translation.
- Redefined goal: To match the closest phrases from the target language
- Albeit thorough matching, *we still deal with longer phrases* (~10 words), an improvement factor mentioned in the base paper.
- Highly semantically different languages.

# ***Importance***

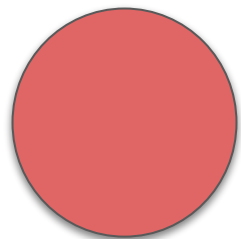
- Effective use of a shared semantic space is crucial for low-resource language translation
- Explores the alignment of heterogeneous modalities (text + video)

# Datasets

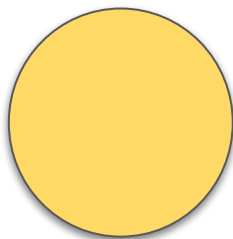
Languages
<i>Chinese (Mandarin)</i>
<i>German</i>

Dataset	Rationale
VATEX ( <i>Chinese - English</i> Video Captions)	For Chinese <i>text-video</i> embeddings
VATEX (Machine Translated <i>English - German</i> )	For German <i>text-video</i> embeddings
VATEX Videos (sans captions)	For <i>video</i> embeddings

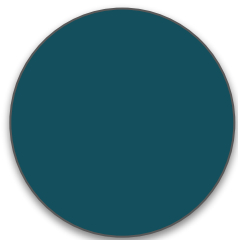
# ***Learning Mappings for Embedding Spaces***



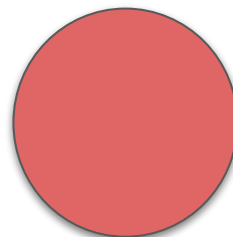
*Chinese*



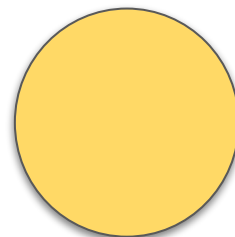
*German*



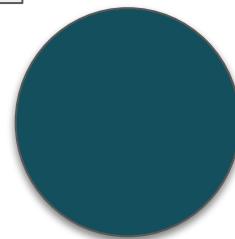
*Video*



***Chinese***



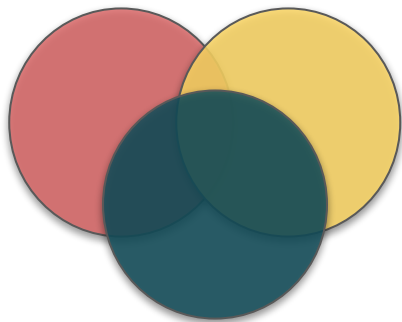
***German***



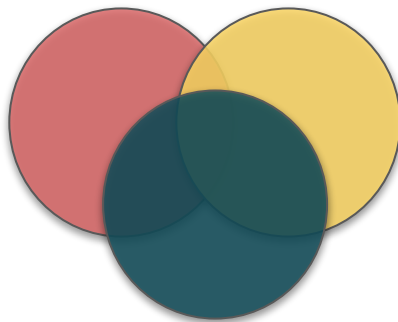
***Video***



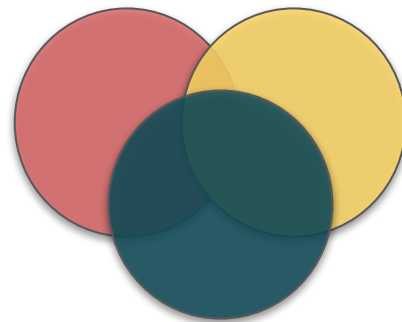
# Transforming the Embedding Space



*Topic 1*



*Topic 2*



*Topic 3*

- Emphasis on **meaning** rather than **modality**
- **Similar** meanings: **Pull** modalities together
- **Dissimilar** meanings: **Push** modalities farther

# Triplet Loss - Base

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

Anchor: a  
reference  
sample (video  
embedding)

Positive: a  
semantically  
similar sample  
(from captions)  
to the anchor

Negative: a  
dissimilar  
sample (caption  
from a different  
video)

L2 Norm is sensitive  
to larger differences -  
good for us!

Margin( $m$ ): a  
hyperparameter

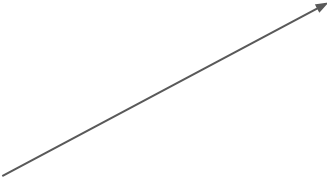
Brings  $a$  and  $p$  closer than  $a$  and  $n$  by at least a margin  $m$



# ***Regularization***

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

Prevents feature  
collapse - keeps  
anchor embedding  
mean away from  
zero

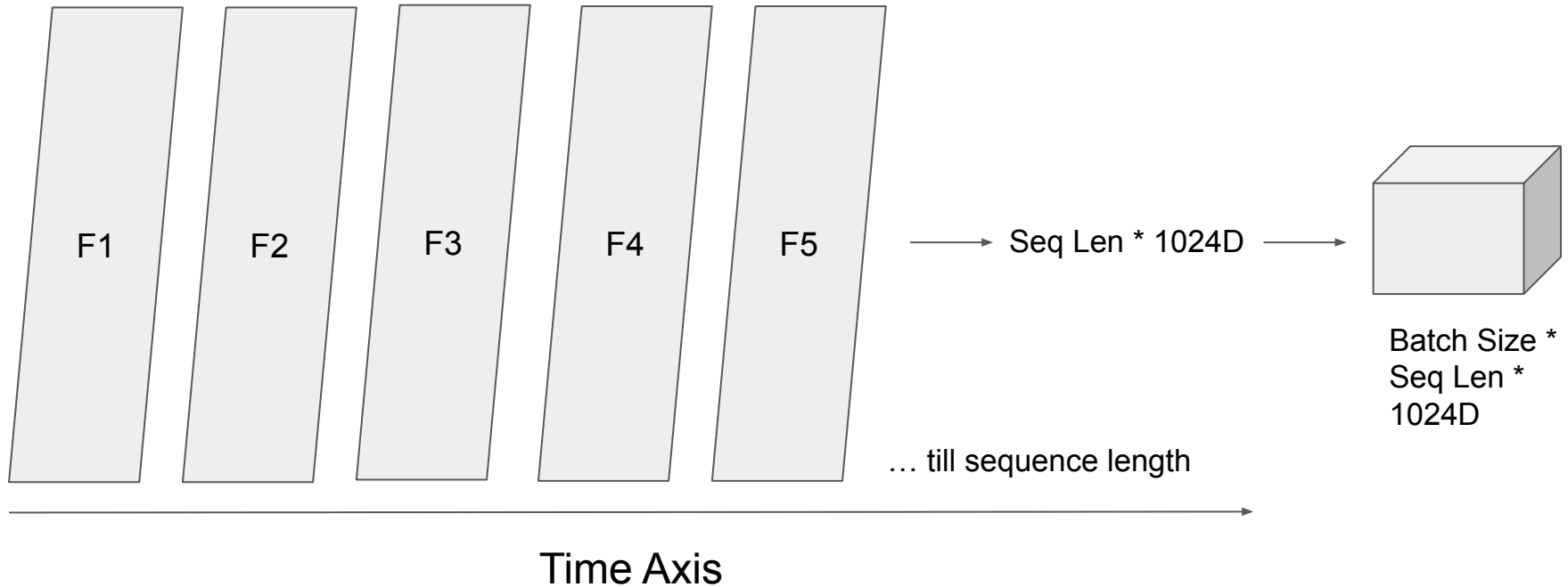


# ***Pairwise Combination of Losses***

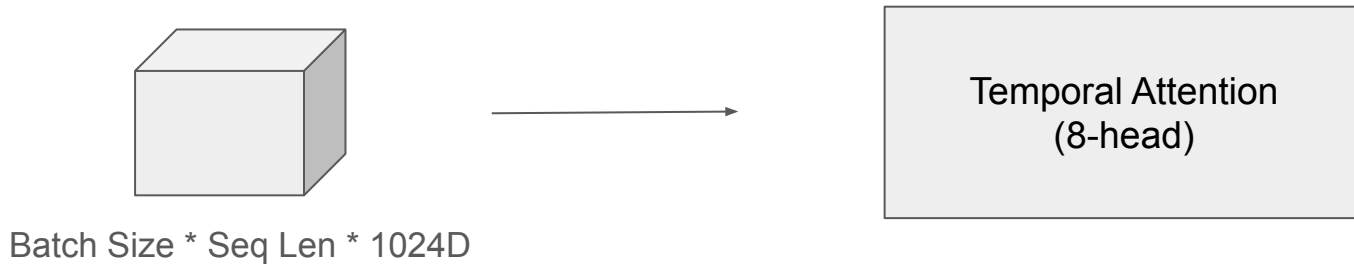
Anchor	Positive	Negative
V	C	$C \rightarrow L_{vc}$
V	G	$G \rightarrow L_{vg}$
C	G	$G \rightarrow L_{cg}$
C	V	$V \rightarrow L_{cv}$
G	V	$V \rightarrow L_{gv}$
G	C	$C \rightarrow L_{gc}$

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

# ***Input Representation - Video***



# ***Temporal Attention Block - Video***



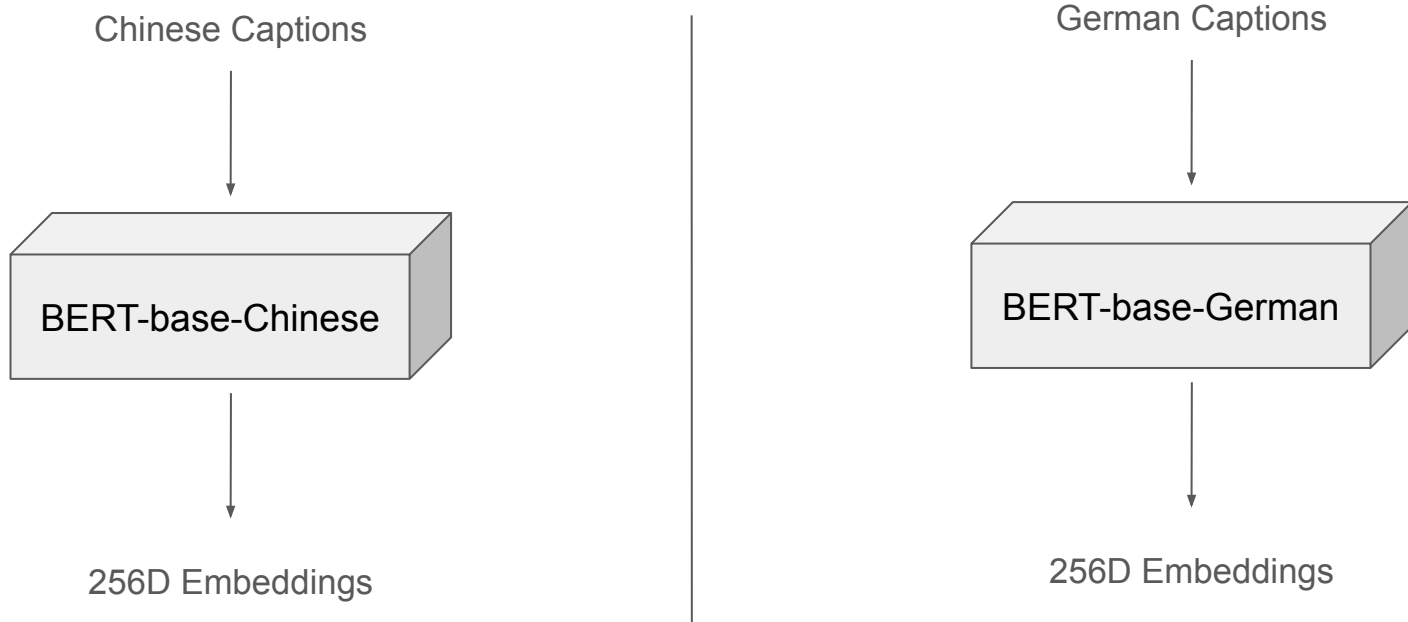
- Processes Inter-Frame Relationships
- Maintains the same dimensions

# Mean Pooling - Video



We've now summarized across frames

# ***Text Processing***



\* with minor additional normalization steps

# ***Experimental Setting (Total)***

Step 1: Inputs (Video Embeddings + Tokenized Chinese, German Captions)

Step 2: Video Embeddings Transformation (Temporal Pooling + MLP), BERT + Projection Layers for Chinese & German

Step 3: (Chinese, German) to Video transformation, along with cross-lingual transform

Step 4: Loss Computation - Pairwise Triplet Loss + Regularization

Step 5: Prediction: Chinese Input => Embed => Match Nearest Video => Match German Captions (top 3)

# Results - *Correct Predictions*

Chinese	Expected German	Predicted German	Similarity
一个人拿着呼啦圈用手在头部转动 随后掉入了腰部。 A person spins a hula hoop with hands on head, then drops it to waist	Eine Dame dreht einen Hula Hope Ring auf ihren Wast. A lady spins a hula hoop on her waist	Eine Dame dreht einen Hula Hope Ring auf ihren Wast. A lady spins a hula hoop on her waist	0.943 - Rank 1
一个小朋友正在一个房子里面玩跑 步机。 A child is playing on a treadmill inside a house	Ein junges Mädchen versucht, ein Treadmill zu verwenden, das sehr dunkel ist. A young girl tries to use a treadmill, it's very dark	Ein junges Mädchen versucht, ein Treadmill zu verwenden, das sehr dunkel ist. A young girl tries to use a treadmill, it's very dark	0.922 - Rank 2



# Results - Misses

Chinese	Expected German	Predicted German	Similarity
一个人拿着呼啦圈用手在头部转动 随后掉入了腰部。 A person spins a hula hoop with hands on head, then drops it to waist	Eine Dame dreht einen Hula Hope Ring auf ihren Wast. A lady spins a hula hoop on her waist	Ein junges Mädchen versucht, ein Treadmill zu verwenden, das sehr dunkel ist. A young girl tries to use a treadmill, it's very dark	0.9223- Rank 2
一个小朋友正在一个房子里面玩跑 步机。 A child is playing on a treadmill inside a house	Ein junges Mädchen versucht, ein Treadmill zu verwenden, das sehr dunkel ist. A young girl tries to use a treadmill, it's very dark	Eine Dame dreht einen Hula Hope Ring auf ihren Wast. A lady spins a hula hoop on her waist	0.9425 - Rank 1

# ***Insights***

- The model is able to retrieve similar phrases, even though the sentences are long (> than 3-4 words, ~ 10 words).
- The expected phrase generally appears in the top 3 predictions.
- Ranking of the phrases seems to be misaligned at times, indicating room for improvement.
- Hyperparameter-tuning (margin adjustment) can be explored to form clearer tri-clusters.

***Thank You!***