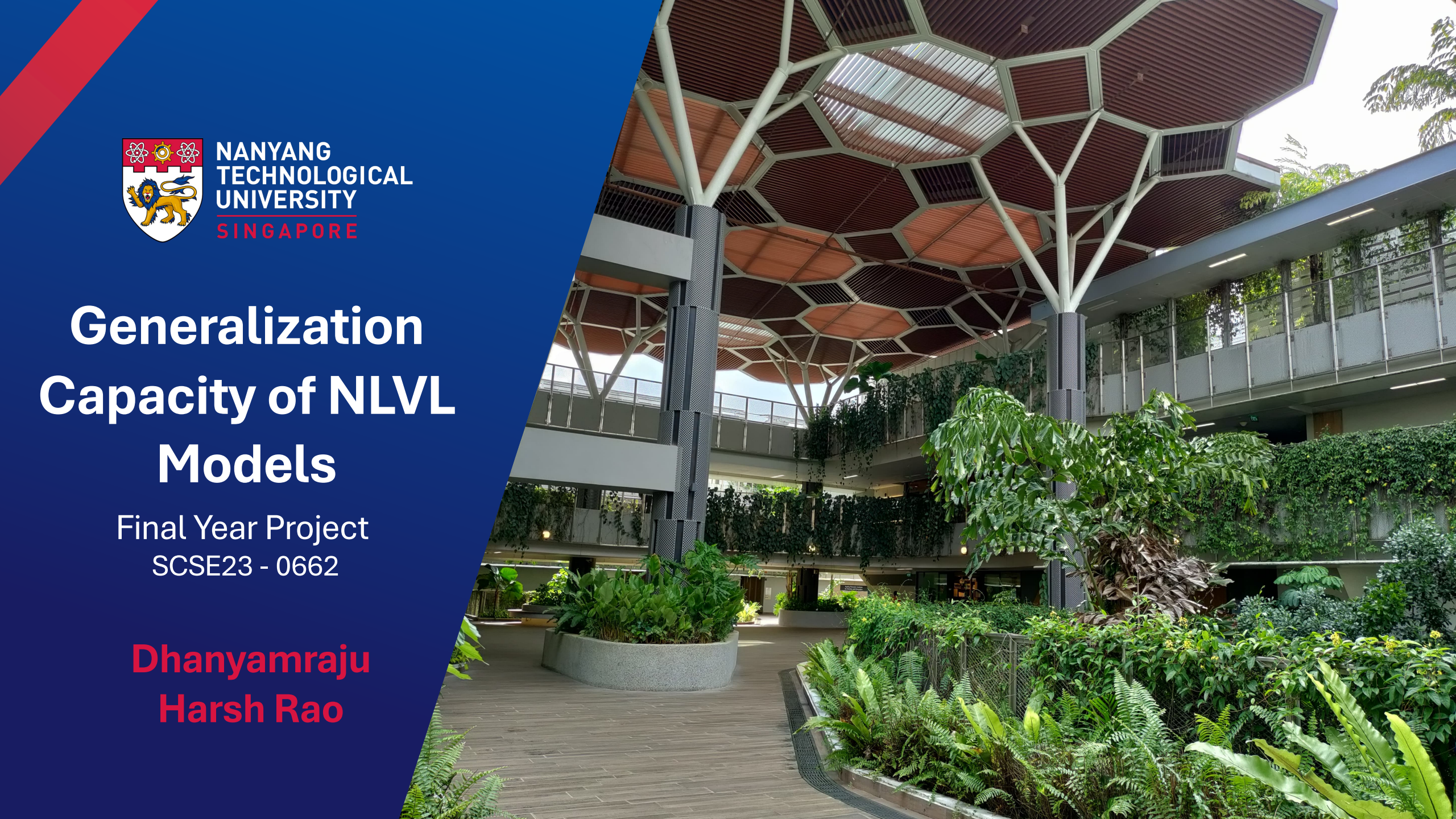


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Generalization Capacity of NLVL Models

Final Year Project
SCSE23 - 0662

Dhanyamraju Harsh Rao

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

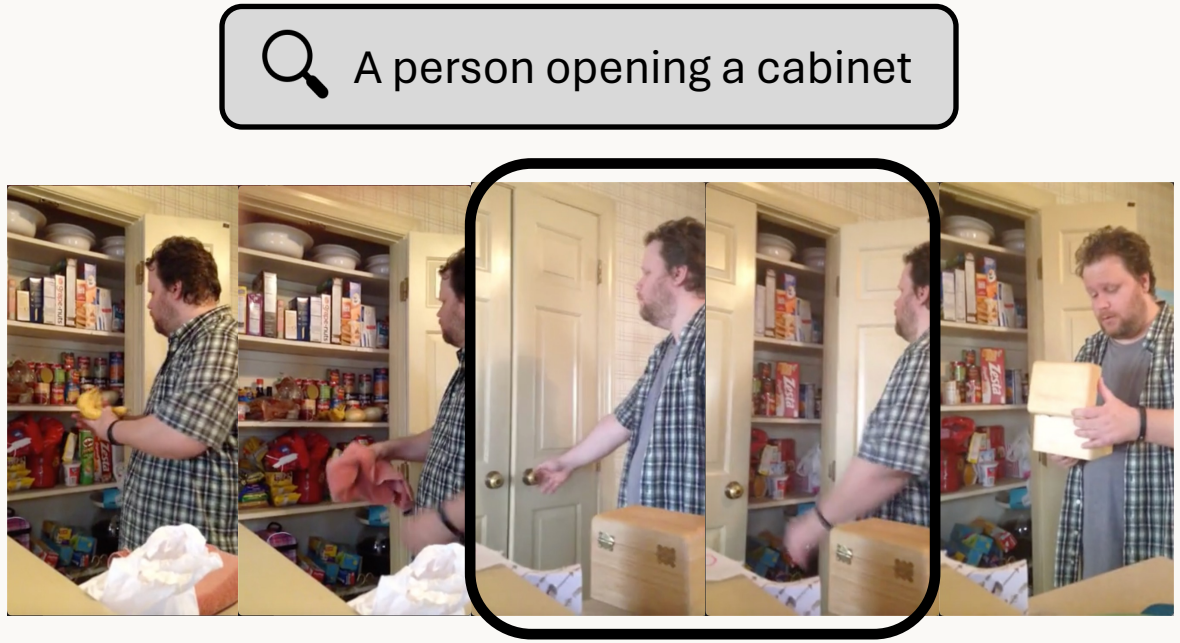
Contents

Background

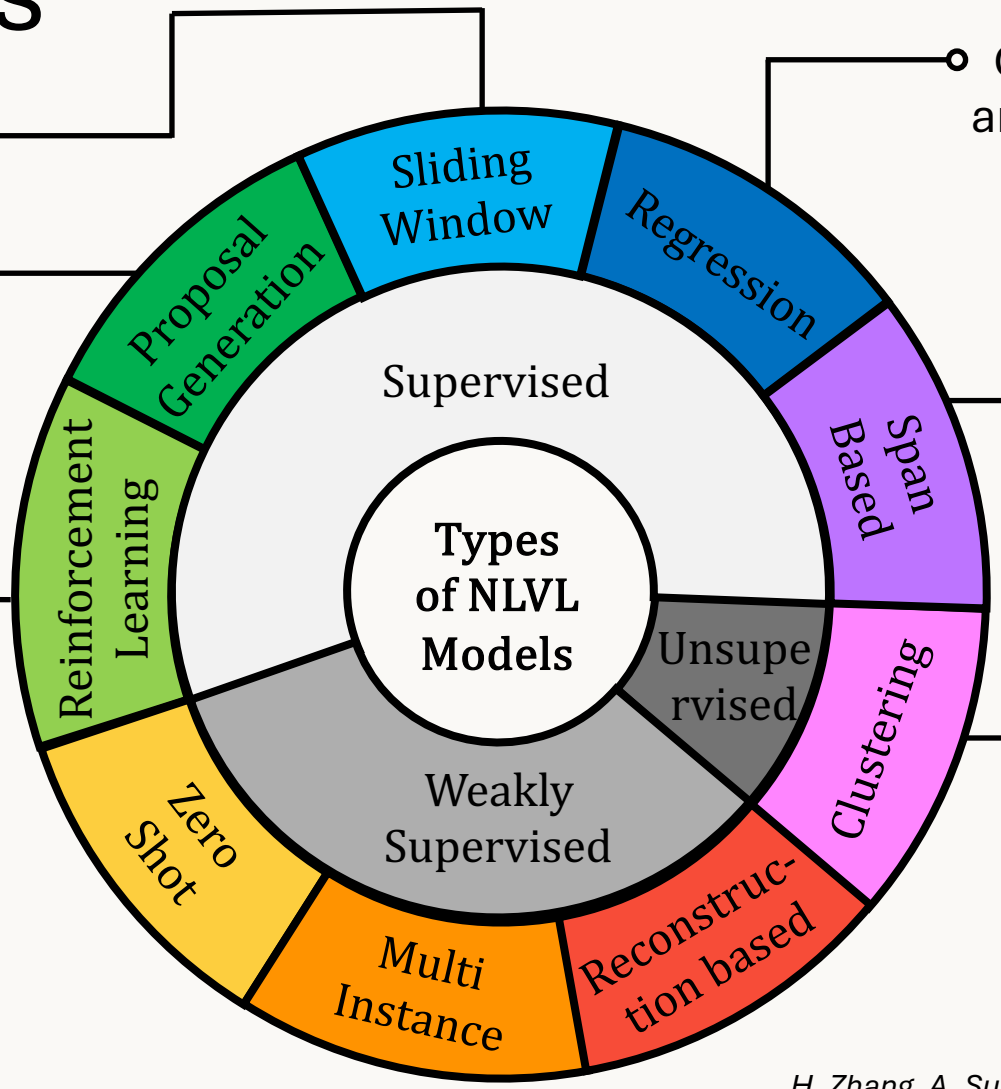
Introduction to NLVL, VSLNet, Charades-STA, and Evaluation Metrics

Introduction to NLVL

- NLVL stands for Natural Language Video Localization
- Its also called Temporal Sentence Grounding in Videos (TSGV) and Video Moment Retrieval (VMR)
- Returns a segment of the video that best matches a text query
- Intersection of Computer Vision, Natural Language Processing, and Information Retrieval
- Applications in Security and Education



NLVL Models



Generates Proposed Solutions using a sliding window and ranks them

Calculates the start and end timestamp of the target segment

Generates Proposed Solutions using model and ranks them

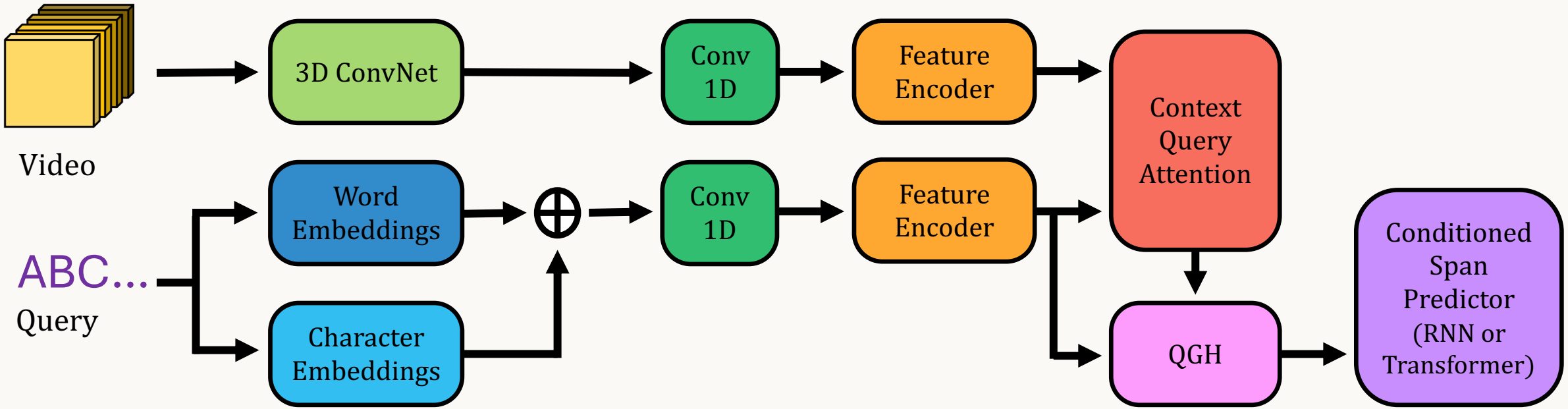
Calculates probability that each frame is the start or end frame of the target segment

Train an agent to adjust the target segment by choosing from certain actions & reward it when it gets closer to the ground truth

Only requires video-query pairs (No Timestamps)

H. Zhang, A. Sun, W. Jing and J. T. Zhou, "Temporal Sentence Grounding in Videos: A Survey and Future Directions," 13 March 2023.

VSLNet (NLVL Model)



- Model used in all experiments
- RNN & Transformer Variants

H. Zhang, A. Sun, W. Jing and J. T. Zhou, "Span-based Localizing Network for Natural Language Video Localization," July 2020.

Charades – STA Dataset

- Created from the Charades Dataset[1]
- People performing regular household tasks
- Mostly shot indoors
- Contains Video-Query Pairs labelled with starting and ending timestamps

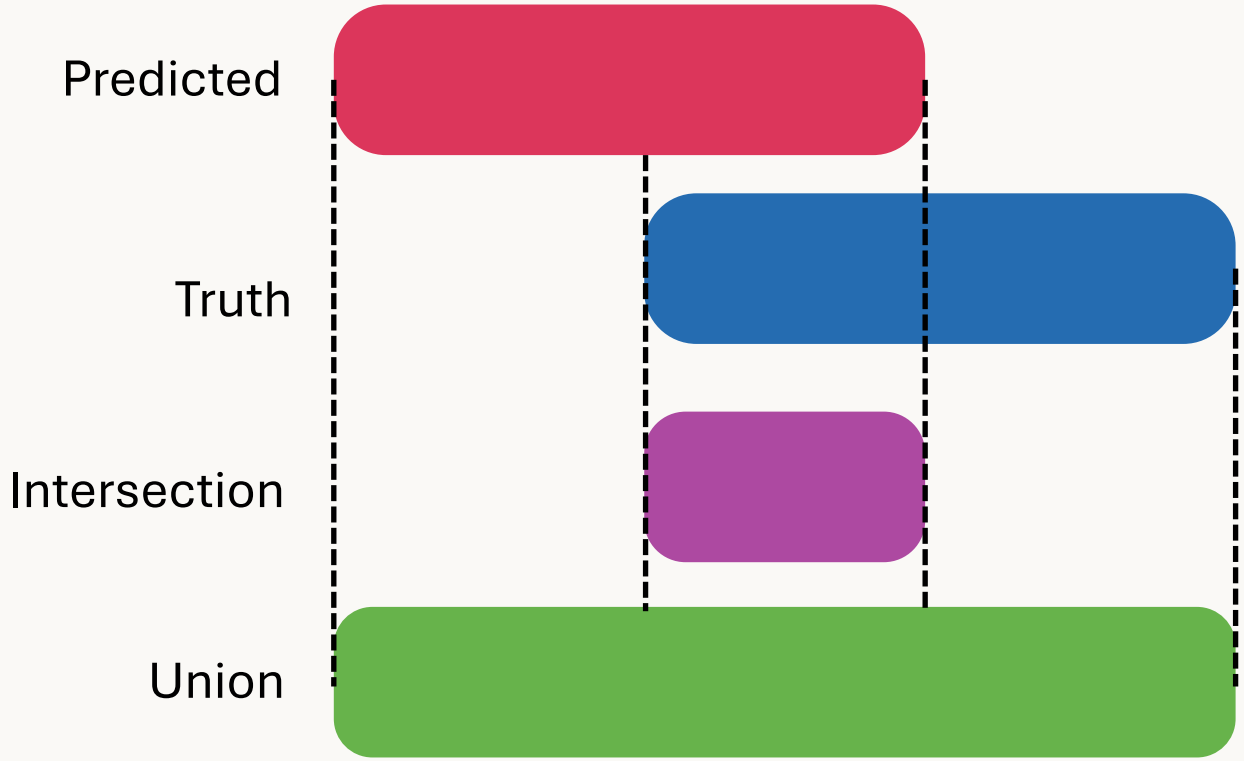
Set	Videos	Annotations
Training	5,338	12,408
Test	1,334	3,720

Charades-STA Dataset[2]

1) *G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev and A. Gupta, "Charades Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding,"*
2) *J. Gao, C. Sun, Z. Yang and R. Nevatia, "TALL: Temporal Activity Localization via Language Query," 3 August 2017*

Evaluation Metrics

- IoU: Intersection over Union
- mIoU: mean IoU
- IoU=m: Percentage of samples with IoU ≥ m
- IoU = 0.3
- IoU = 0.5
- IoU = 0.7



$$IoU = \frac{min(e_{pred}, e_{truth}) - max(s_{pred}, s_{truth})}{max(e_{pred}, e_{truth}) - min(s_{pred}, s_{truth})} \times 100$$

Experiment – 1

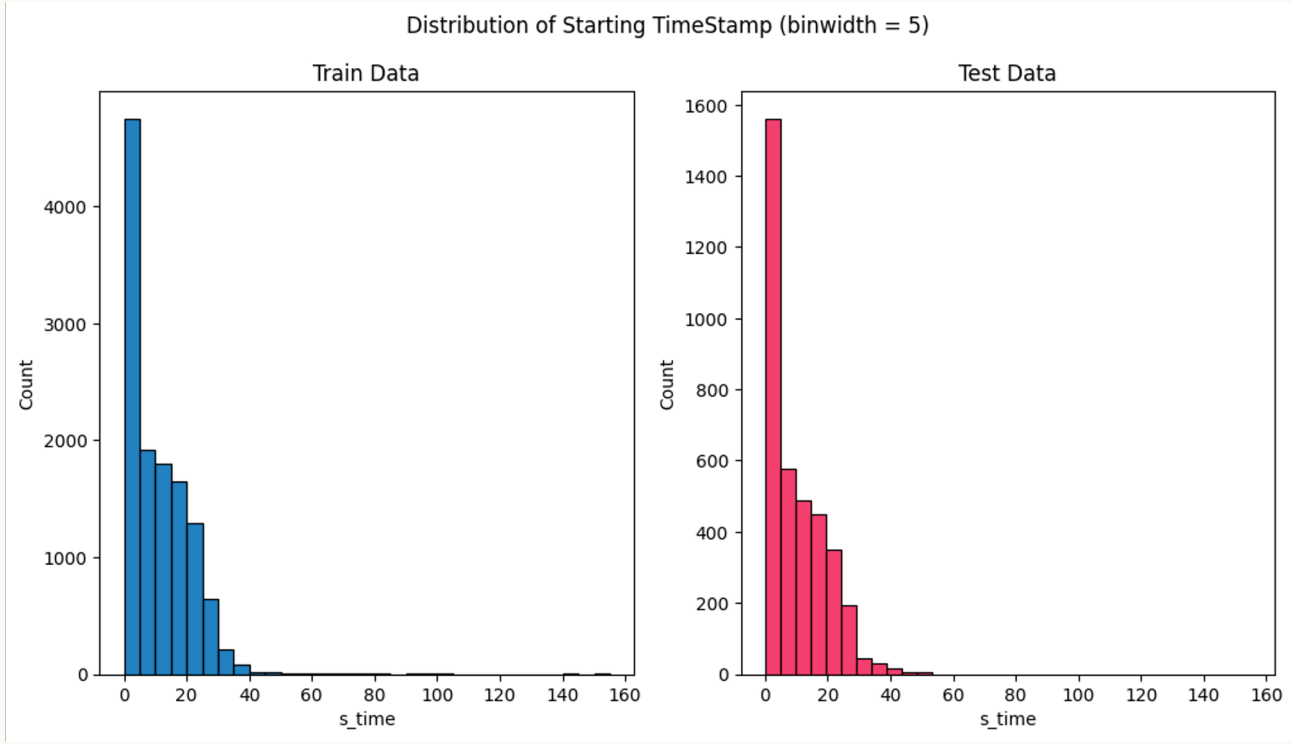
Charades – STA - Merged

Removing temporal bias from the Charades-STA dataset

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Motivation – Distribution Bias

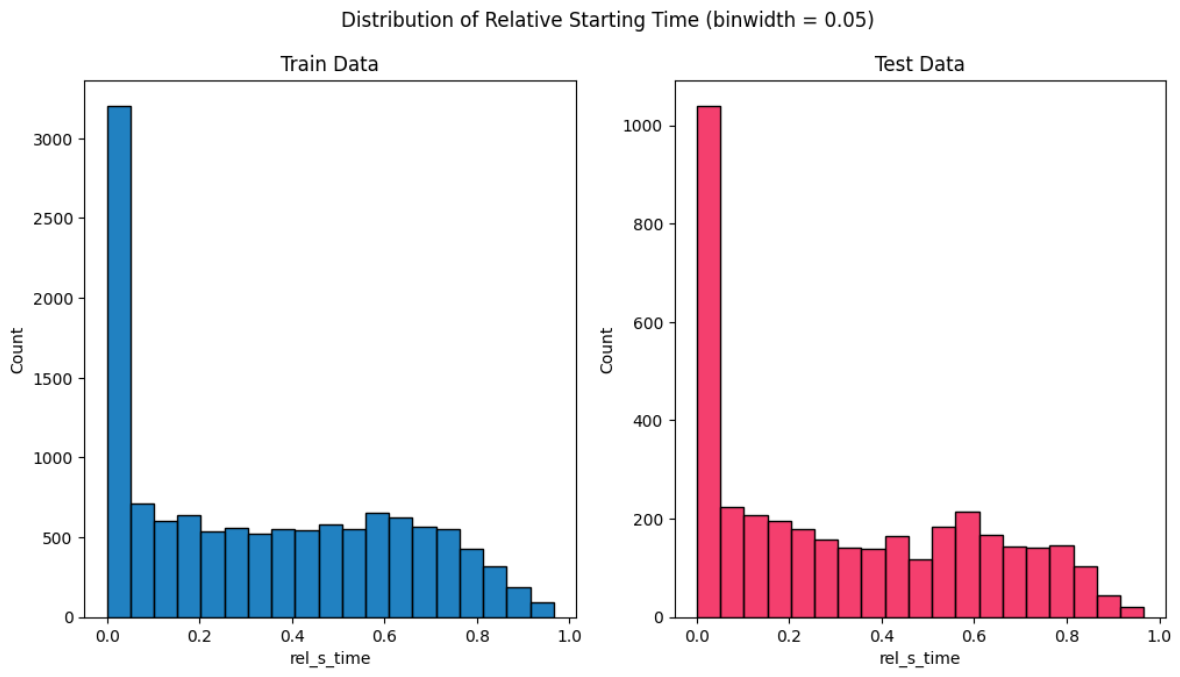
- Temporal Distributional Bias in Charades-STA
- Strong skew towards 0
- Consistent in training and test data



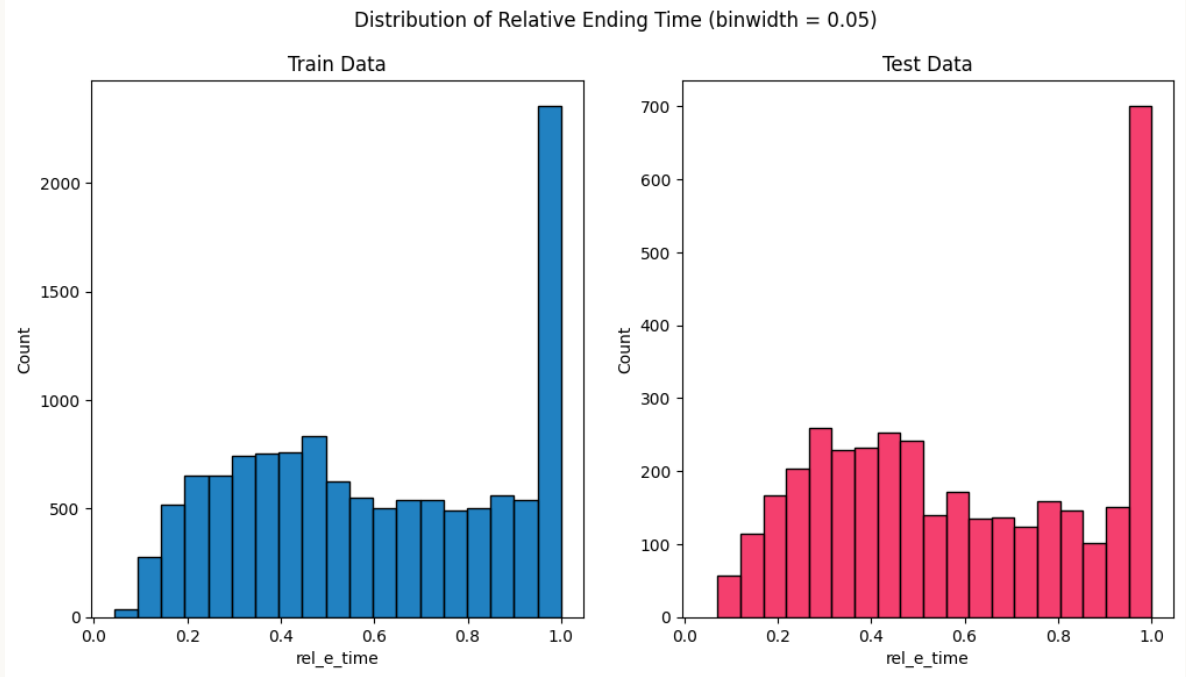
Starting Timestamp Distribution

Motivation – Distribution Bias

- Relative Start Time = Start Time / Video Duration



Relative Starting Timestamp Distribution



Relative Ending Timestamp Distribution

Motivation – Bias Biased Models

1) Full Video – Always returns the entire video
2) ConstTime3 – Returns a constant Start & End timestamp
   i. Optimized to Maximise IoU = 0.3
   ii. On Training Data
   iii. Nelder-Mead Optimizer

Model	Start Time	End Time
Full Video	0.0	Video Duration
ConstTime3	0.1633	16.1169

Model	Year	IoU = 0.3	IoU = 0.5
MCN	2017	13.57	4.05
ROLE	2018	25.26	12.12
Full Video	2024	35	0.43
ConstTime3	2024	51.69	18.68
QSPN	2019	54.7	35.6
VSLNet	2020	70.46	54.19
CPN	2021	75.73	59.77
URL	2022	77.88	55.69

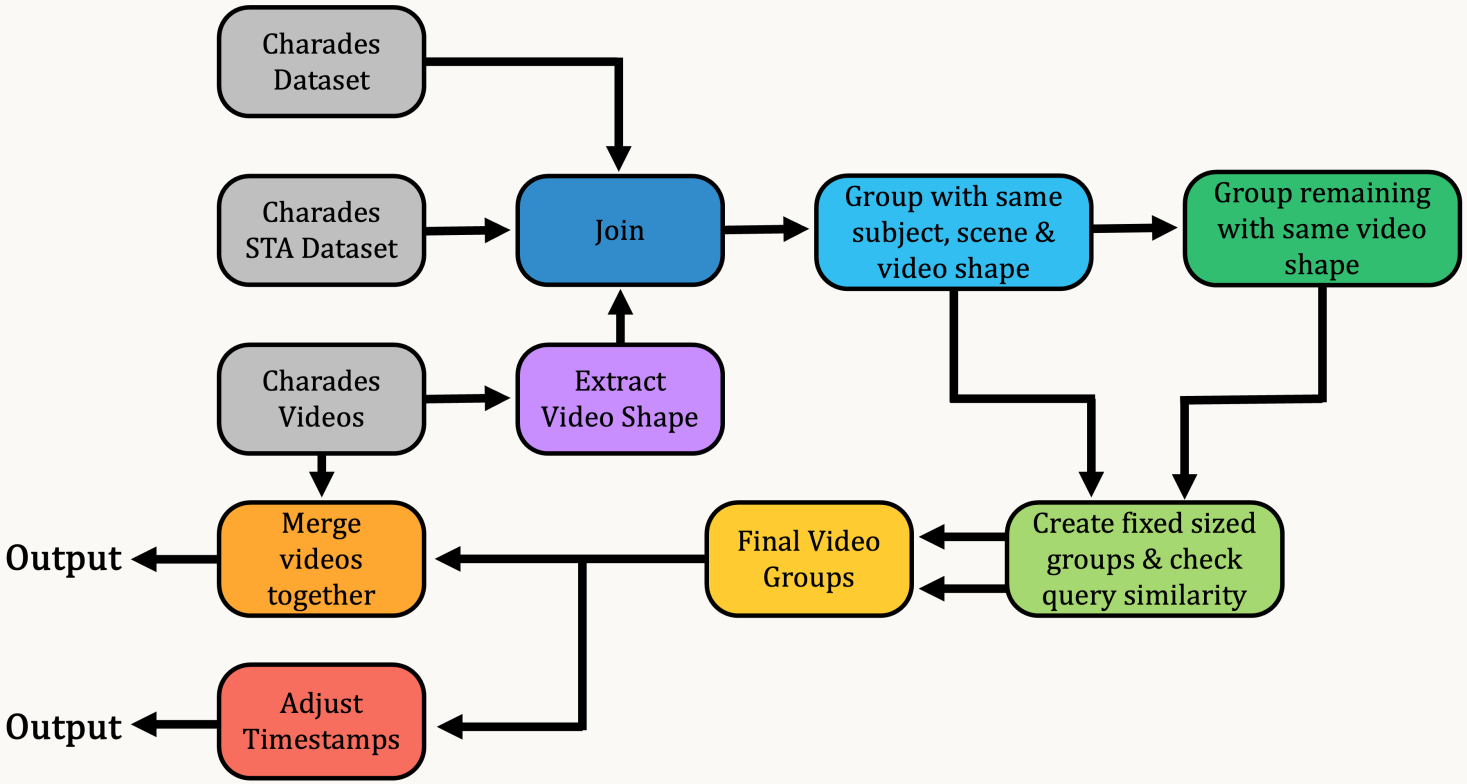
F. Gao and L. Han, "Implementing the Nelder-Mead simplex algorithm with adaptive parameters," January 2010

Objective

- Reduce Temporal bias in Charades-STA Dataset
- Create a new dataset Charades-STA-Merged by merging different videos together
- Redistribute the timestamps
- Challenges of merging:
  - Same width, height, and frame rate
  - No Similar Queries between videos (can’t introduce new solutions)

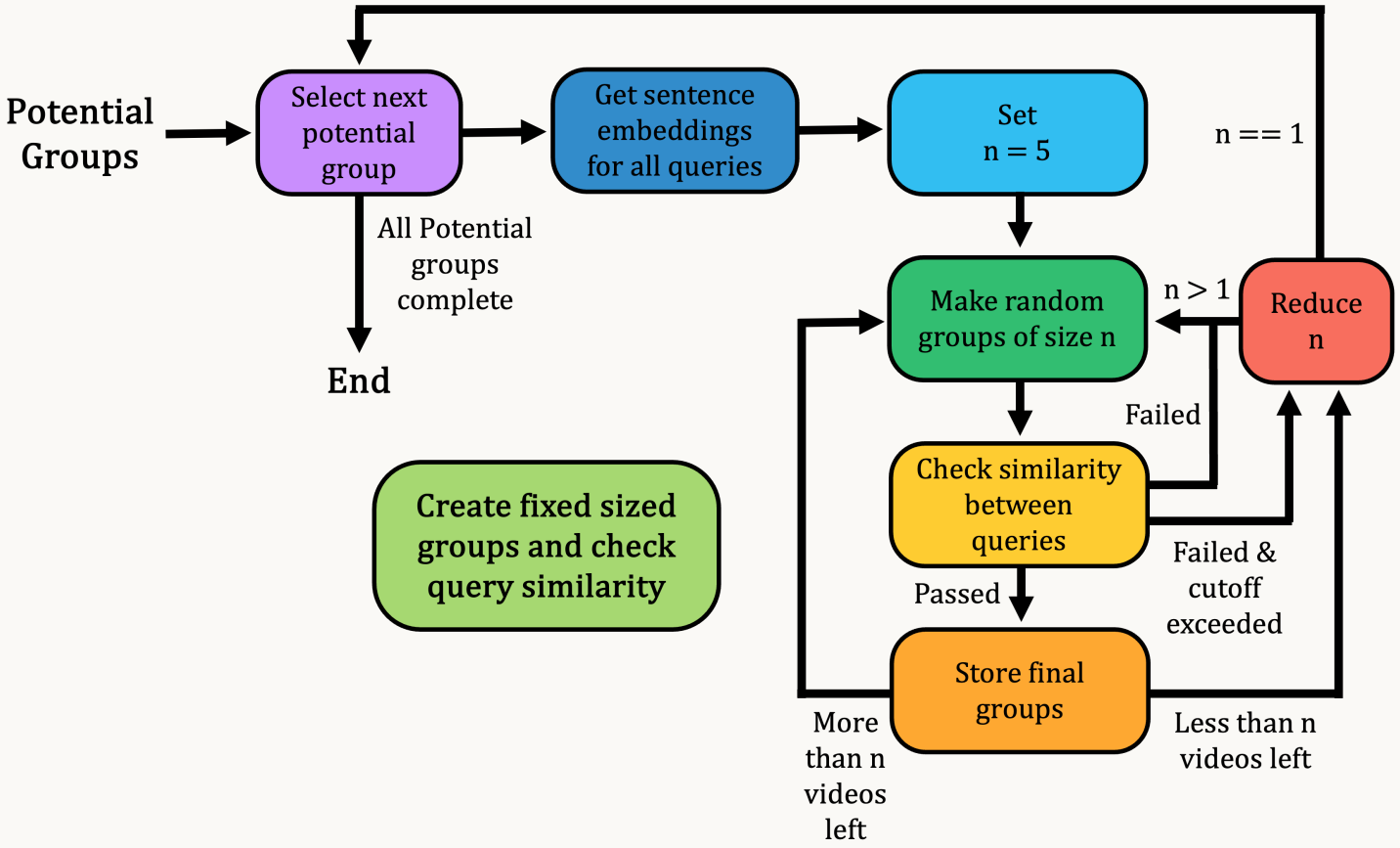
Methodology

- Reduce is temporal distributional bias in the dataset
- Do so by merging videos together to scatter the starting and ending timestamps
- Take care that videos have same width, height, and frame rate
- Take care not to introduce valid timestamps for given video-query pairs



Methodology

- Reduce is temporal distributional bias in the dataset
- Do so by merging videos together to scatter the starting and ending timestamps
- Take care that videos have same width, height, and frame rate
- Take care not to introduce valid timestamps for given video-query pairs



Sentence Transformers, "MiniLM-L6-v2,"

Results

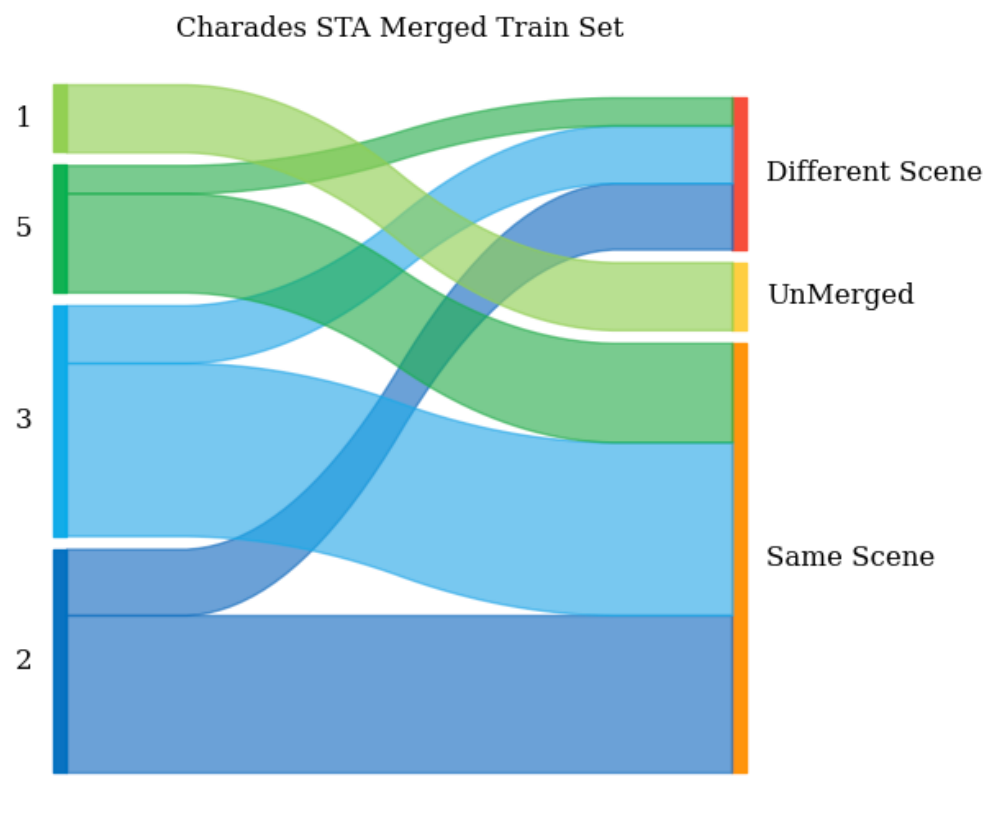
Same Scene (Merged Video Id: LSFJG_Z6B6S_T9Y1N)

LSFJG | Z6B6S | T9Y1N

Different Scene (Merged Video Id: PIJRH_JVOM3)

PIJRH | JVOM3

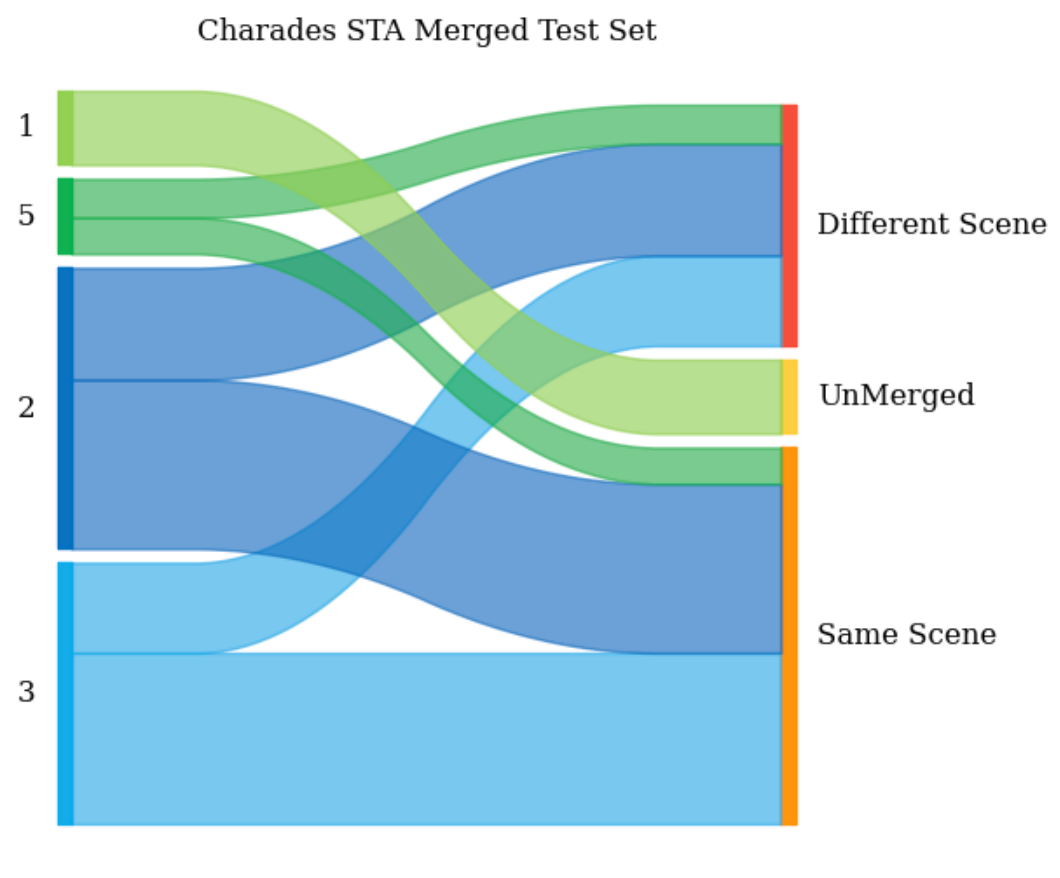
Results



Video-Query pairs in Charades-STA-Merged Train Set

	Same Scene	Different Scene	Total
5 Videos Merged	1595	477	2072
3 Videos Merged	3050	1069	4119
2 Videos Merged	3079	1464	4543
Total Merged	7724	3010	10734
UnMerged	-	-	1671

Results

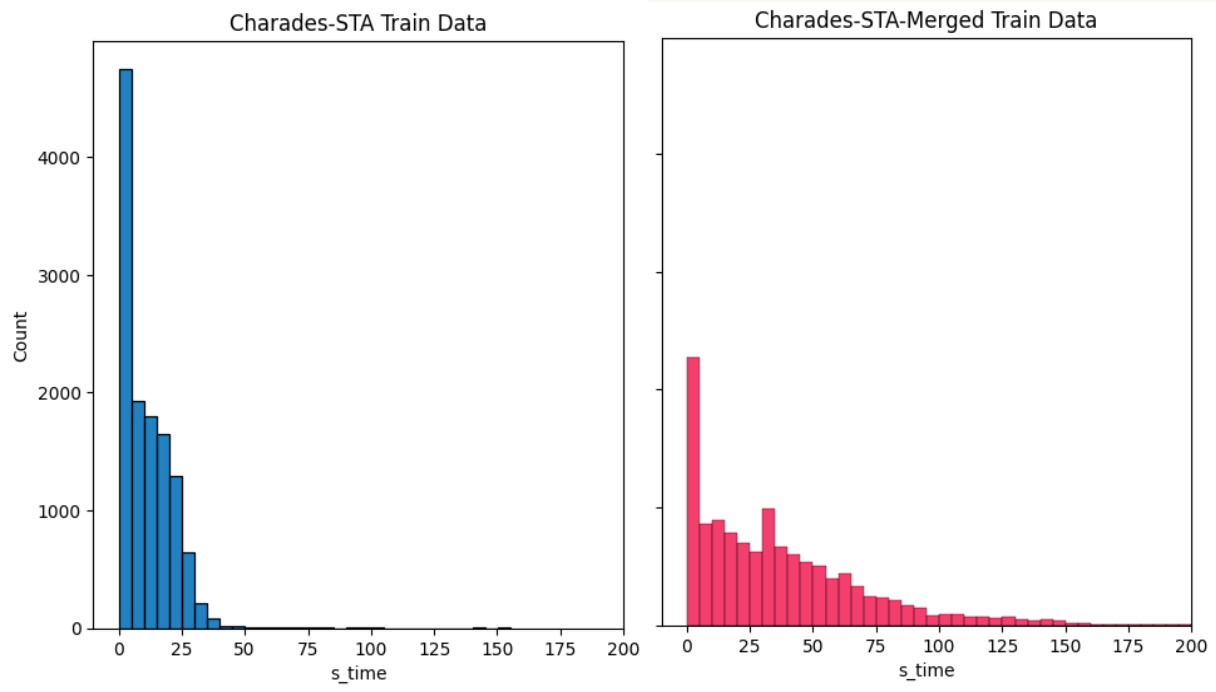


Video-Query pairs in Charades-STA-Merged Test Set

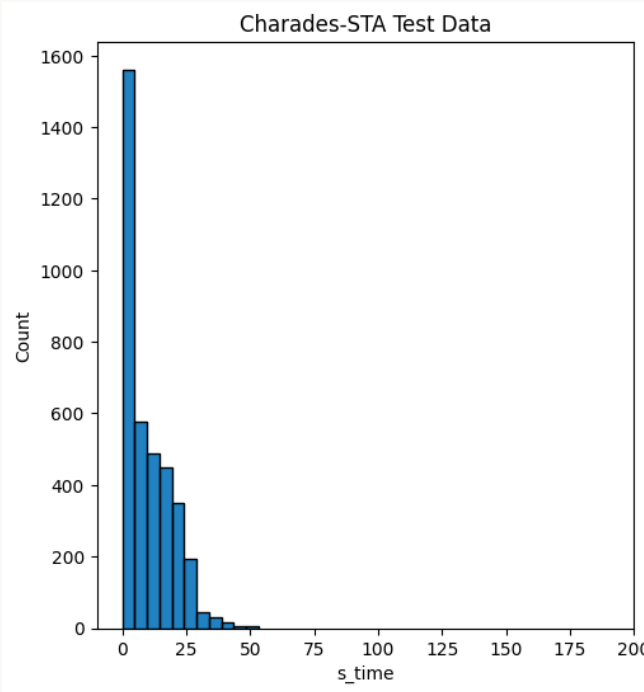
	Same Scene	Different Scene	Total
5 Videos Merged	146	175	321
3 Videos Merged	782	513	1295
2 Videos Merged	866	682	1548
Total Merged	1794	1370	3164
UnMerged	-	-	556

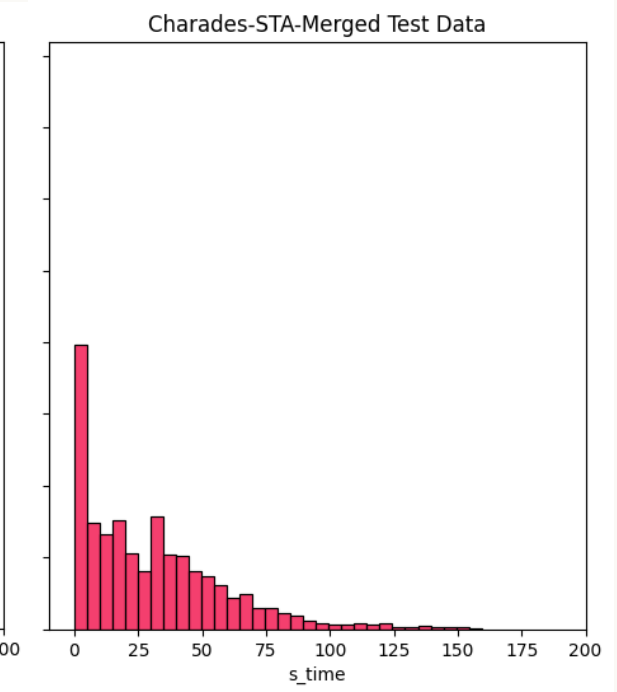
Results

Starting Timestamp Distribution



Train Set



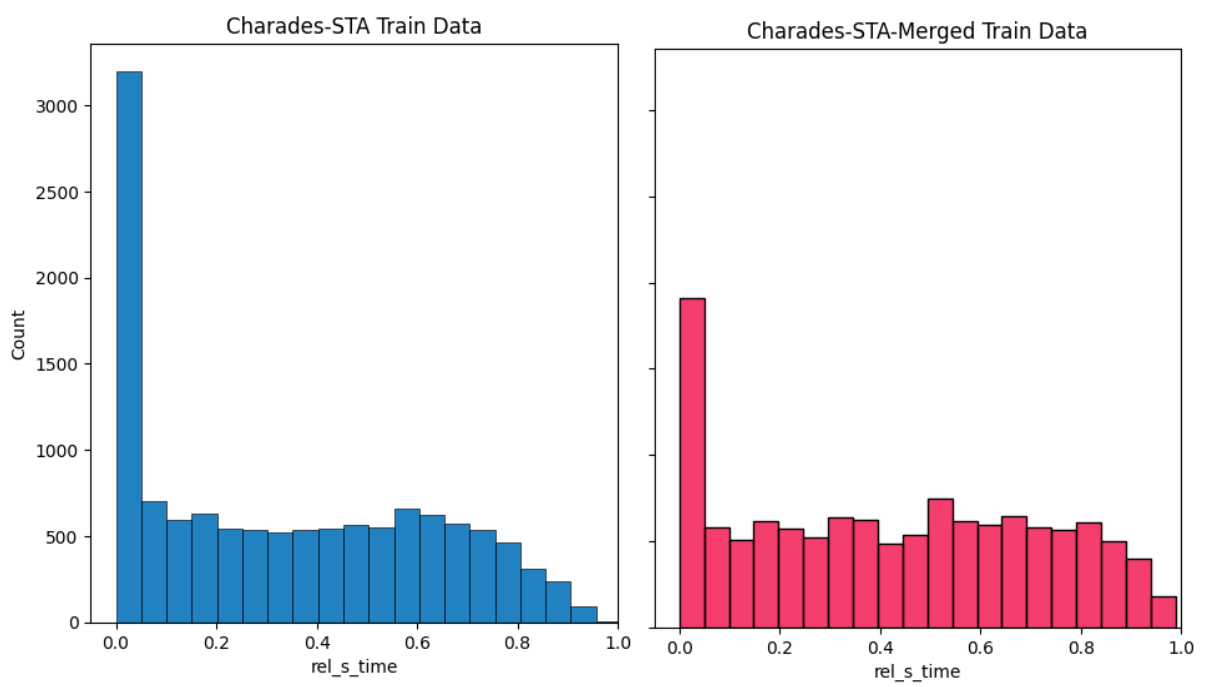


Test Set

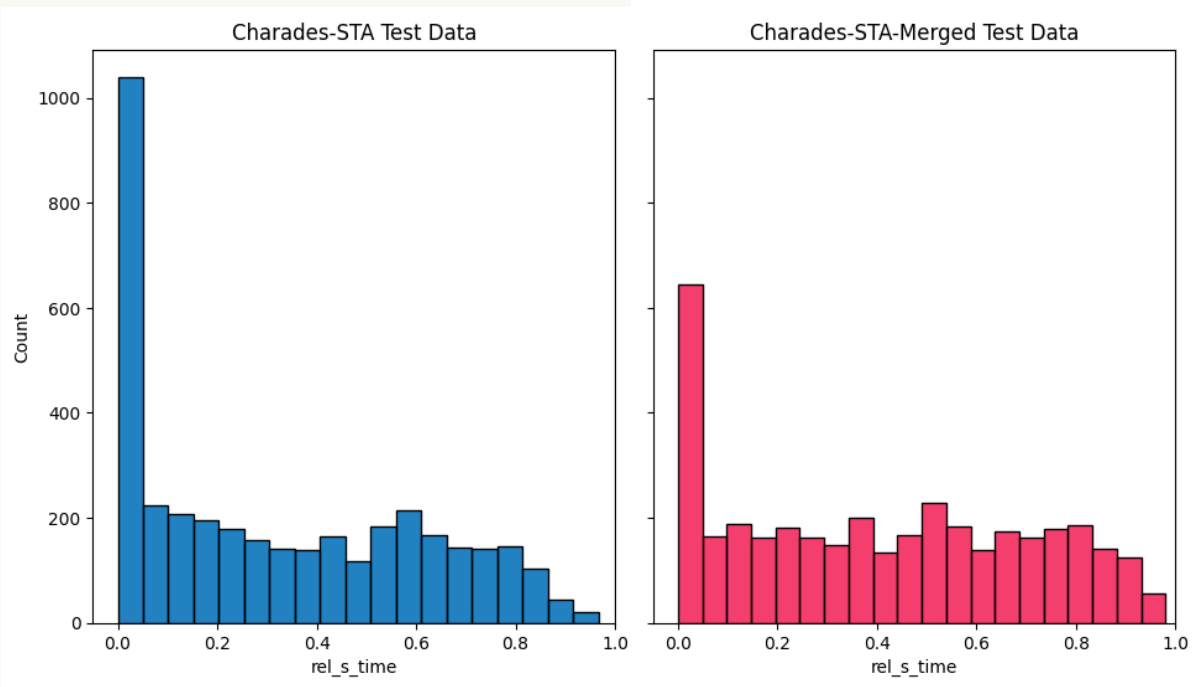
A few samples exist outside the range of the Charades-STA-Merged Train Graph

Results

Relative Starting Time Distribution



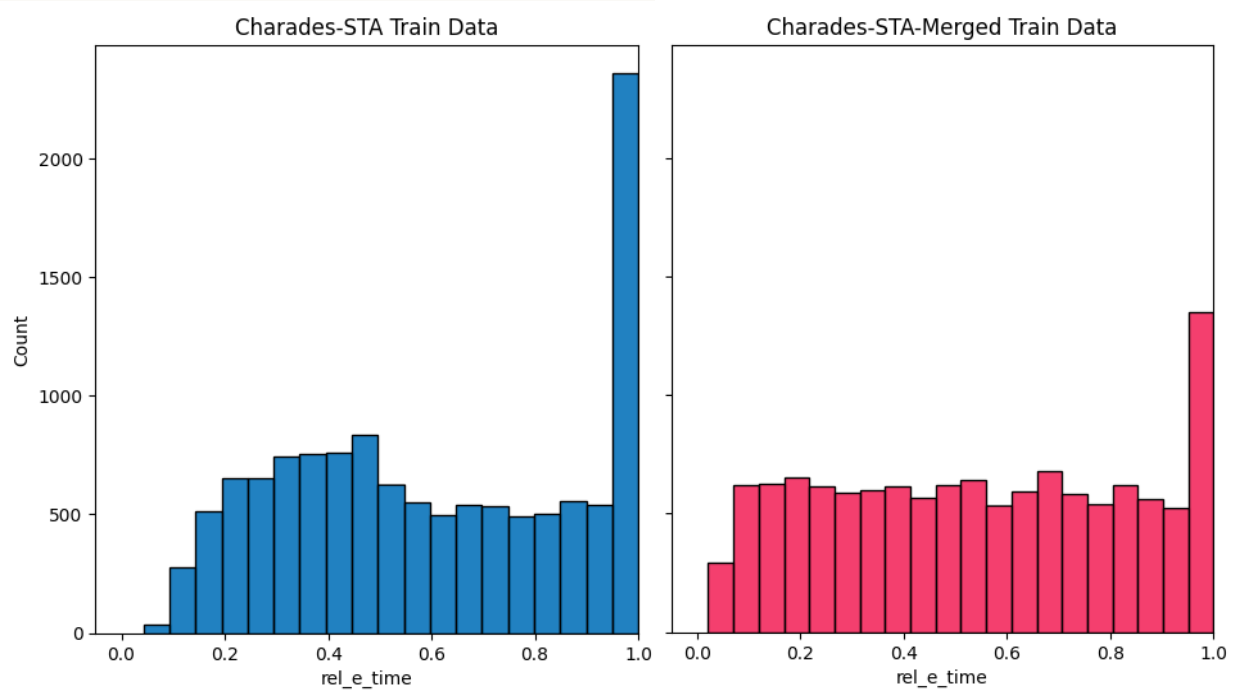
Train Set



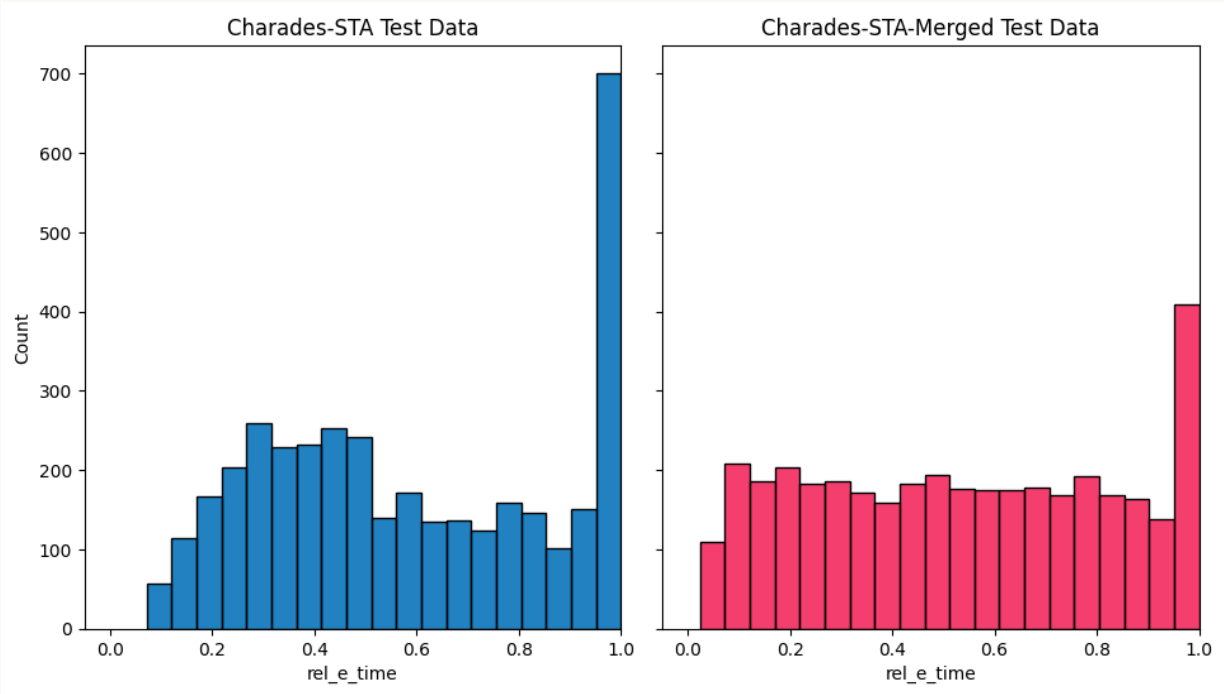
Test Set

Results

Relative Ending Time Distribution



Train Set



Test Set

Results

Bias Based Models

Model	Trained Dataset	IoU = 0.3	IoU = 0.5
FullVideo	Charades-STA	35	0.43
	Charades-STA-Merged	4.4 ↓	0.02 ↓
ConstTime3	Charades-STA	51.69	18.68
	Charades-STA-Merged	14.89 ↓	5.11 ↓

Charades-STA

Model	Start Time	End Time
Full Video	0.0	Video Duration
ConstTime3	0.1633	16.1169

Charades-STA-Merged

Model	Start Time	End Time
Full Video	0.0	Video Duration
ConstTime3	31.38	47.71

Results

VSLNet with Charades-STA-Merged

Model	Dataset Trained On	Dataset Tested On	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU
VSLNet (RNN)	Charades-STA	Charades-STA-Merged	35.91	23.2	11.42	24.88
	Charades-STA-Merged	Charades-STA-Merged	26.37 ↓	15.51 ↓	7.1 ↓	18.7 ↓
VSLNet (Transformer)	Charades-STA	Charades-STA-Merged	36.61	24.14	12.8	25.68
	Charades-STA-Merged	Charades-STA-Merged	25.54 ↓	14.76 ↓	7.37 ↓	19.1 ↓

Despite being trained on a dataset with less bias the model seems to have acquired a greater degree of bias

VSLNet's performance decreases on long videos. Must try with VSLNet-L to verify.

Y. Zeng, D. Cao, S. Lu and H. Zhang, "Moment is Important: Language-Based Video Moment Retrieval via Adversarial Learning," 2022.

Experiment – 2

Charades – Ego - STA

Testing the generalization capabilities towards first person (egocentric) videos

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Motivation & Objective

- Test NLVL Models on First Person Video
- Increasing popularity of Go-Pros and body cams
- Create the Charades-Ego-STA dataset from the Charades-Ego Dataset[1]

First Person Video[2]

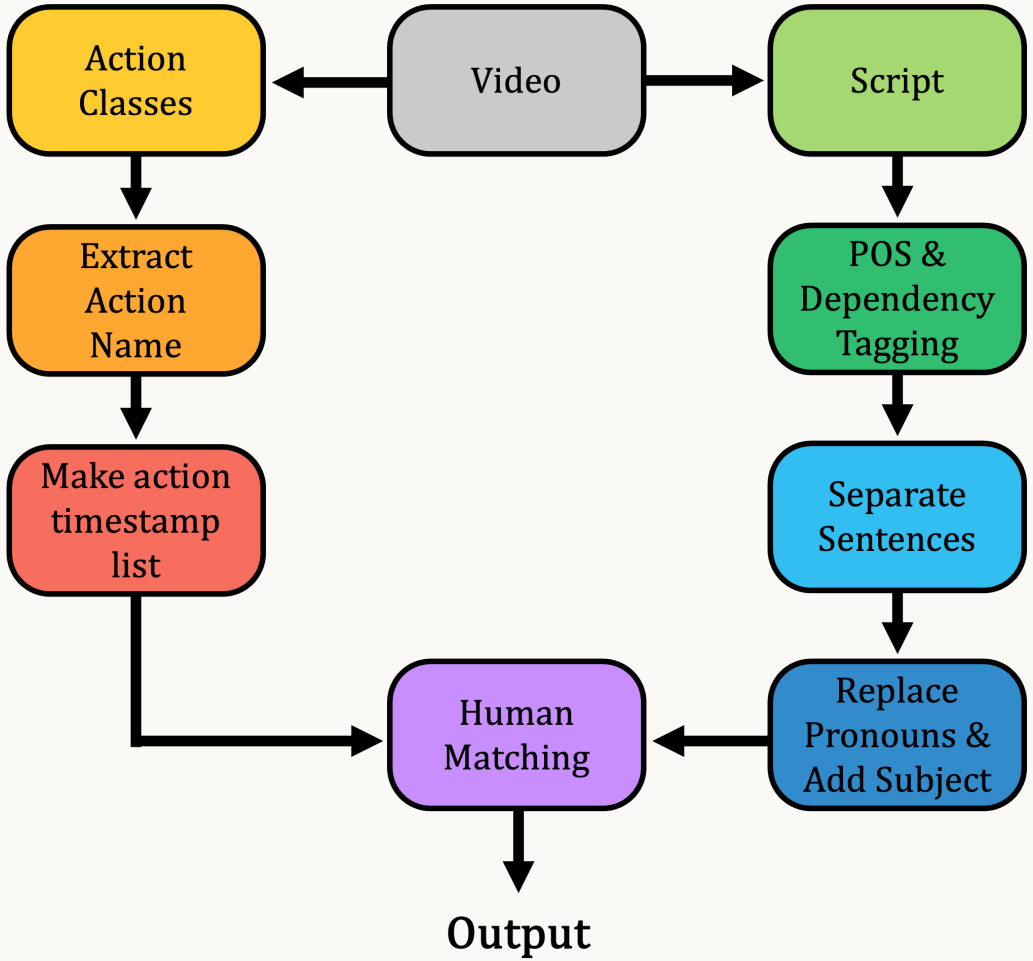
1) G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi and K. Alahari, "Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos," 25 April 2018.

2) https://gizmodo.com.au/2014/06/extreme-biker-rides-through-the-narrowest-mountain-edge-ive-ever-seen/

Methodology

- Create Queries from the given video script
- Manually verify the timestamp – video - query matching

Script	Queries
A person opens the door to their bedroom, then takes shoes off, and holds them up in the air.	A person opens the door to their bedroom.
	(A person) takes their shoes off.
	(A person) holds (them / door / shoes) up in the air.



Results

- Decrease in performance as expected
- Performance Loss similar to Charades-STA-Merged
- Potential for transfer learning
- Need more labelled data

Dataset	Videos	Query-Video Pairs
Charades-Ego-STA	102	302

Model	Dataset Trained On	Dataset Tested On	IoU = 0.3	IoU = 0.5	IoU = 0.7	mIoU
VSLNet (RNN)	Charades-STA	Charades-STA	71.45	54.57	35.27	50.44
	Charades-STA	Charades-Ego-STA-	38.41	21.19	9.27	25.87
	Charades-STA	Charades-STA-Merged	35.91	23.2	11.42	24.88

Experiment – 3

Charades – STA Text Perturbed

Analyzing the generalization capabilities against text perturbations

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Motivation & Objective

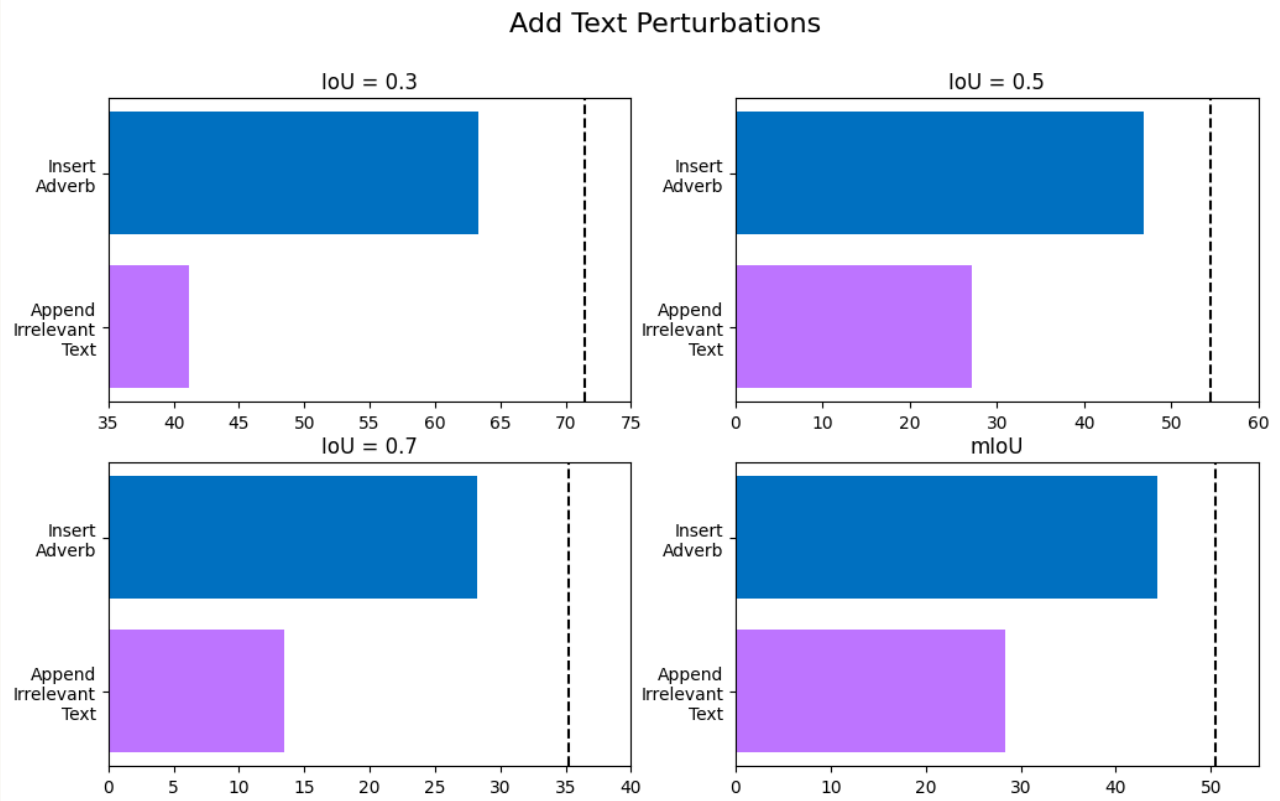
- Real World Data is more messy than clean datasets
- Handling errors in data is very important
- To assess the generalization capabilities of NLVL Models with respect to errors in the query by introducing perturbations
- Used a procedure similar to that used by Schiappa et al.
- All experiments run on VSLNet (RNN) trained on Charades-STA

M. C. Schiappa, S. Vyas, H. Palangi, Y. S. Rawat and V. Vineet, "Robustness Analysis of Video-Language Models Against Visual and Language Perturbations," 18 July 2023.

Add Text Perturbations

- Made using Textflint library
- Model doesn't depend much on adverbs and will not get confused by them easily
- Model struggles to deal with long queries with irrelevant text

Set	Original Query	Perturbed Query
Insert Adverb	Person turn a light on.	Person continuously turn a light on.
Add Irrelevant Text	Person turn a light on.	Besides, person turn a light on.

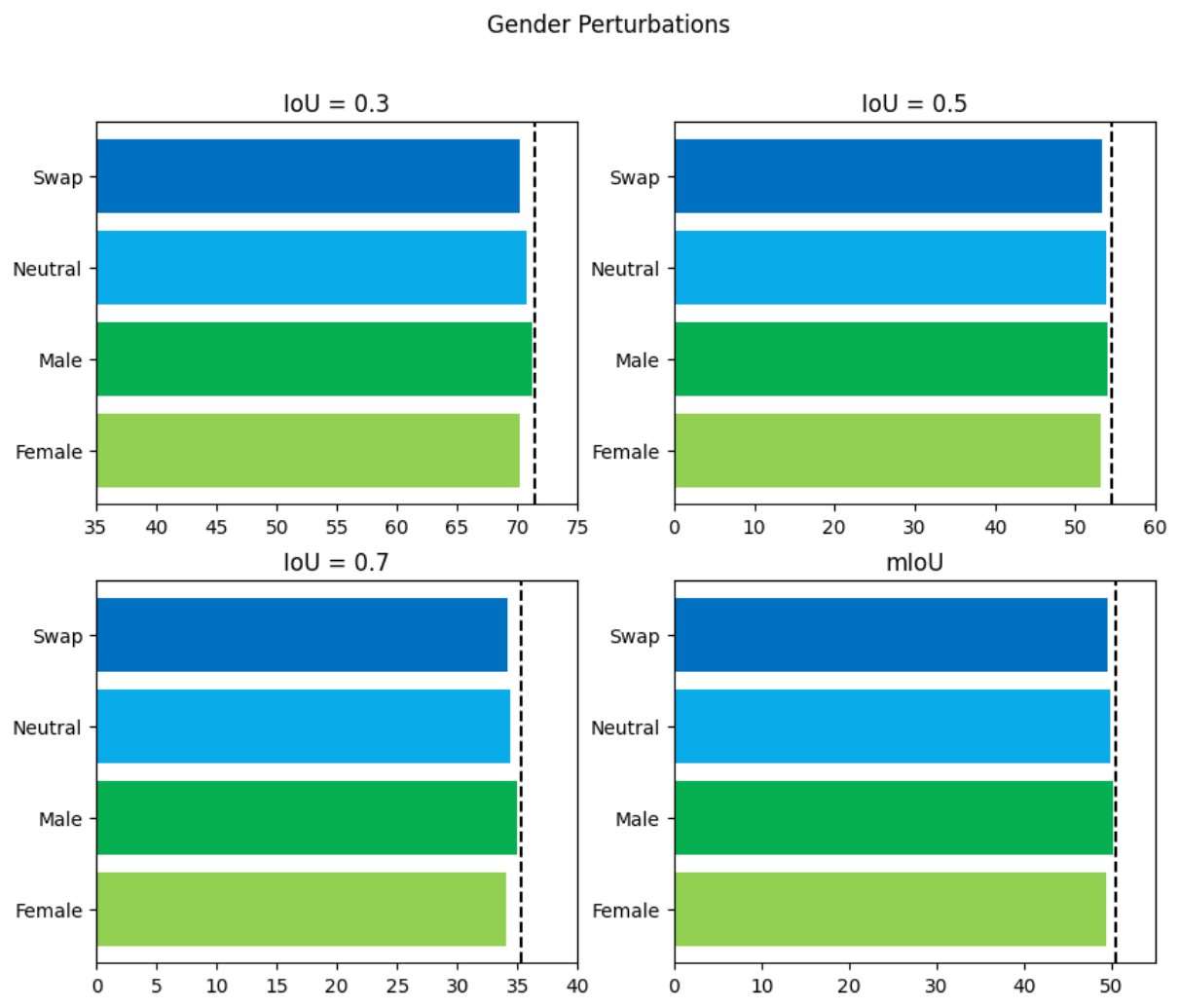


T. Gui, X. Wang, Q. Zhang, Q. Liu, Y. Zou, X. Zhou, R. Zheng, C. Zhang, Q. Wu, J. Ye, Z. Pang, Y. Zhang, Z. Li, R. Ma, Z. Fei, R. Cai, J. Zhao, X. Hu, Z. Yan and Y. Tan, "TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing," 21 March 2021

Gender Perturbations

- Made using gender bender library
- Model is relatively unaffected by any change in gender in query

Set	Original Query	Perturbed Query
Female	The same person was laughing while he was undressing.	The same person was laughing while she was undressing.
Male	She runs out of the room.	He runs out of the room.
Swap	She runs out of the room.	He runs out of the room.
Neutral	She runs out of the room.	*They run out of the room.*



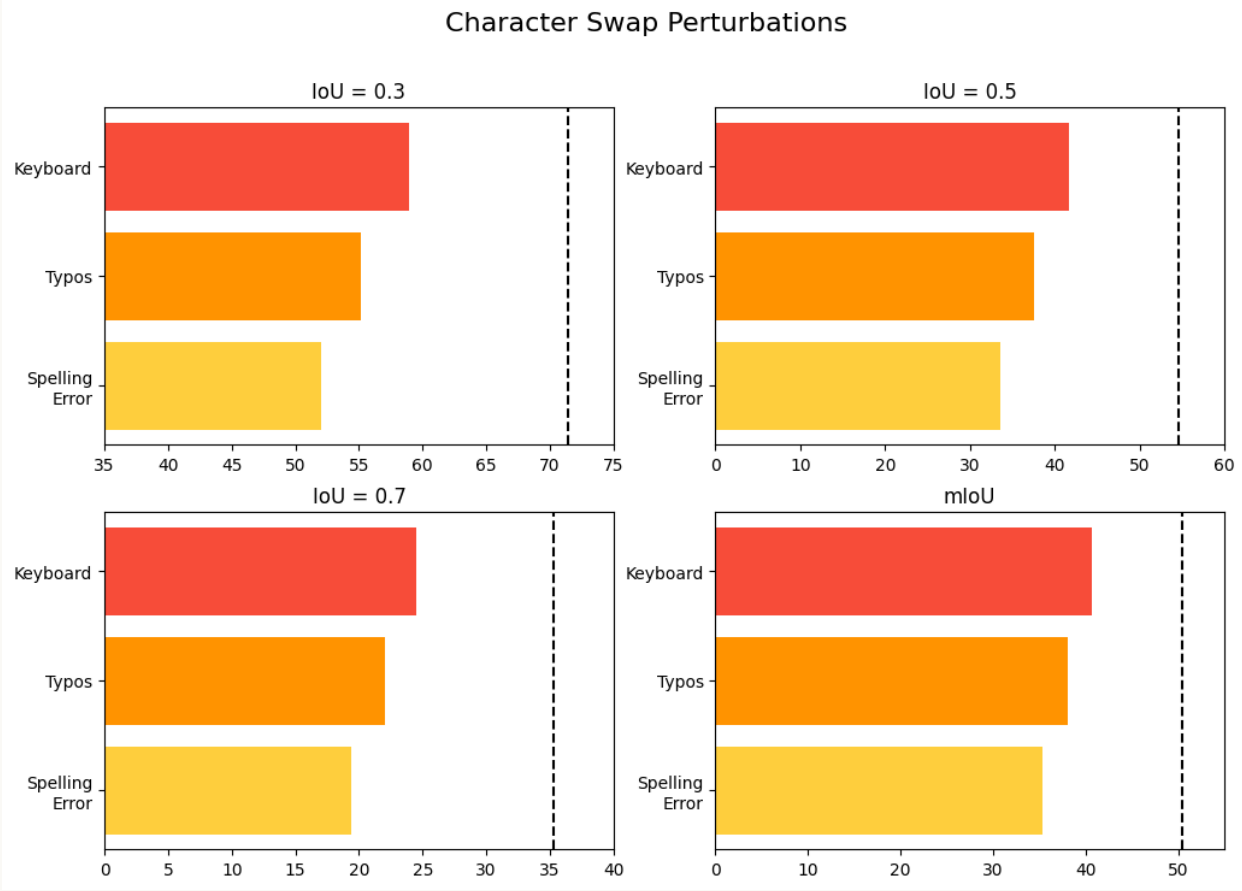
G. Reynolds and J. Wilk, "Gender Bender," 2020

Character Swap Perturbations

- Simulates user mistakes
- All character swaps cause a drop in performance
- Spelling mistakes is the worst as it can cause confusion in the embedding layer

Set	Original Query	Perturbed Query
Keyboard	Person closing the door.	Person closing the coor.
Typos	Person turn a light on.	Person tun a light on.
Spelling Error	A person sits on a chair.	A person sites on a chear.

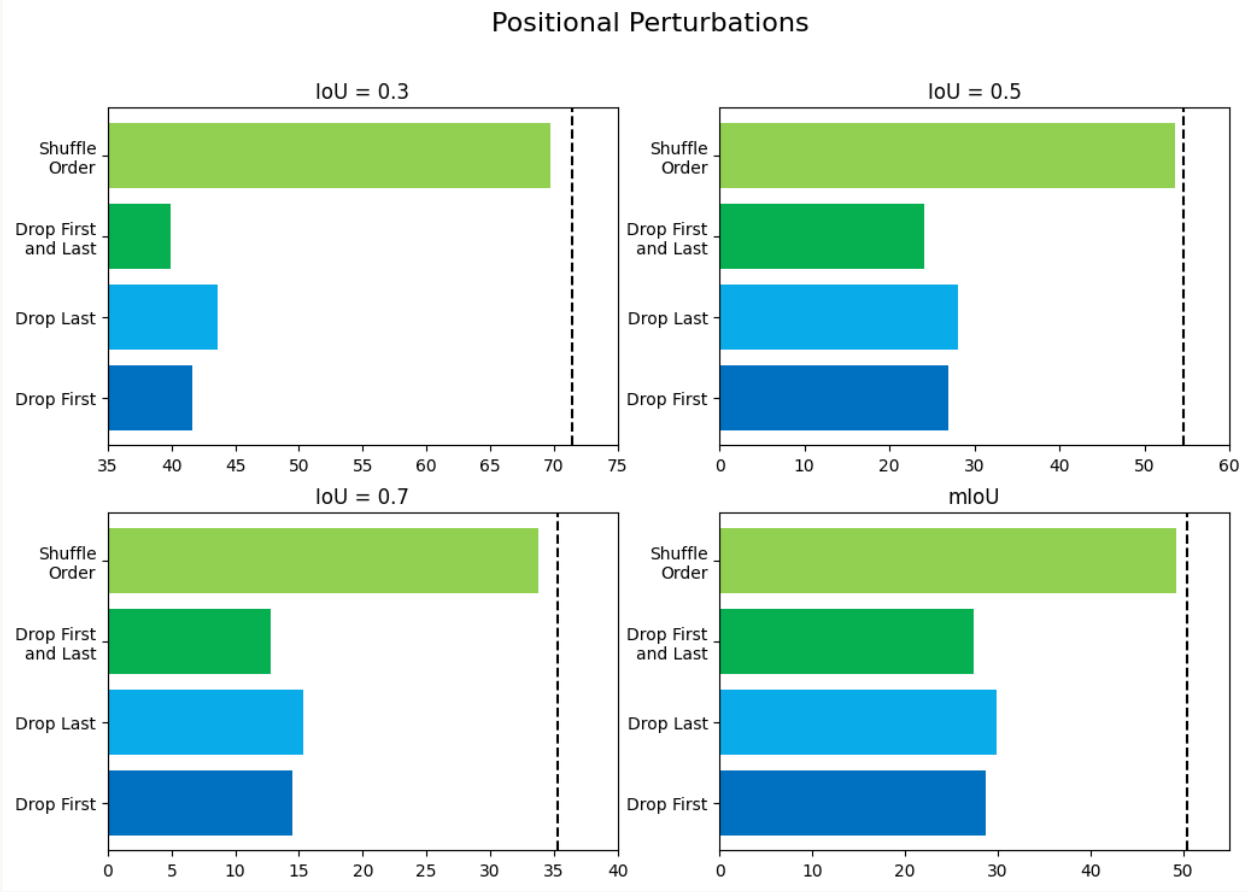
‘chear’ = ‘chair’ or ‘cheer’ ?



Positional Perturbations

- Model doesn't make use of order of words in the query
- First and Last Word have a high chance of being the subject and object which explains the poor performance

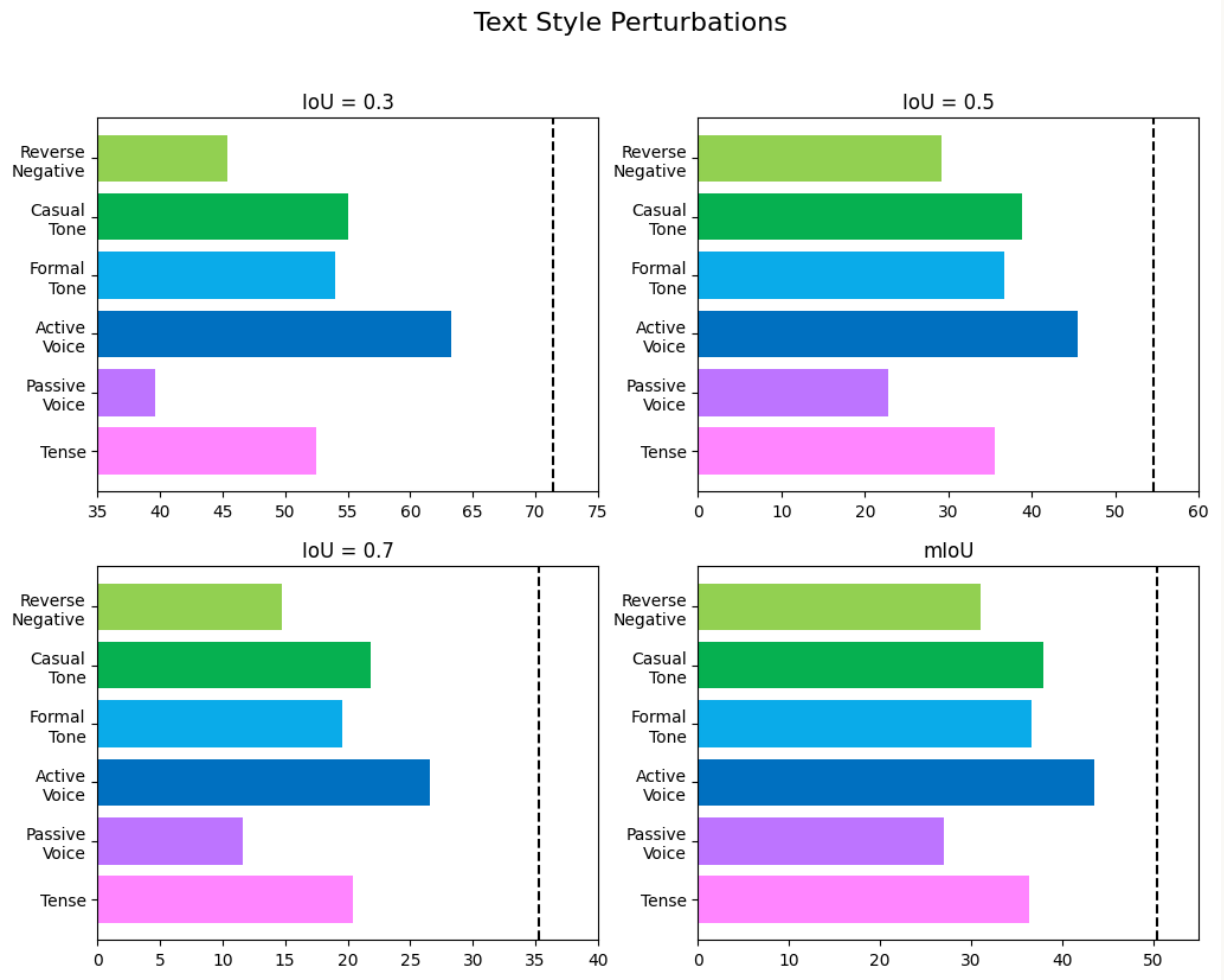
Set	Original Query	Perturbed Query
Shuffle Order	*Person turn a light on.*	*a on. light person turn*
Drop First	*Person turn a light on.*	*<UNK> turn a light on.*
Drop Last	*Person turn a light on.*	*Person turn a light <UNK>.*
Drop First & Last	*Person turn a light on*	*<UNK> turn a light <UNK>.*



Text Style Perturbations

- Used the StyleFormer Library

Set	Original Query	Perturbed Query
Reverse Negative	Person turn a light on.	Person does not turn a light on.
Casual Tone	The same person was laughing while he was undressing.	The person laughing was undressing.
Formal Tone	The same person was laughing while he was undressing.	The same person was laughing while he was undressing.
Active Voice	A dish is washed by a person.	Person washes a dish.
Passive Voice	Person turn a light on.	A light is turned on by a person.
Tense	Person turn a light on.	Person turned a light on.

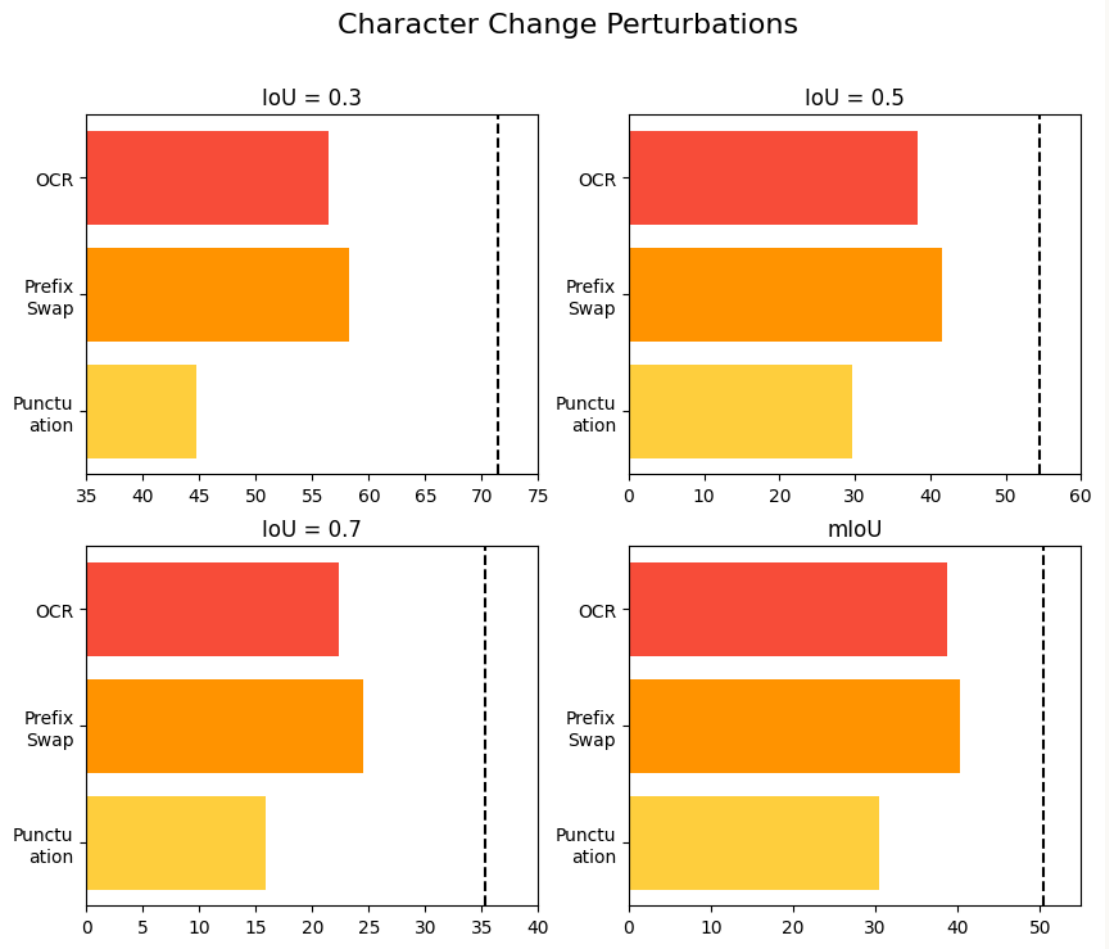


P. Damodaran, "Styleformer,"

Character Change Perturbations

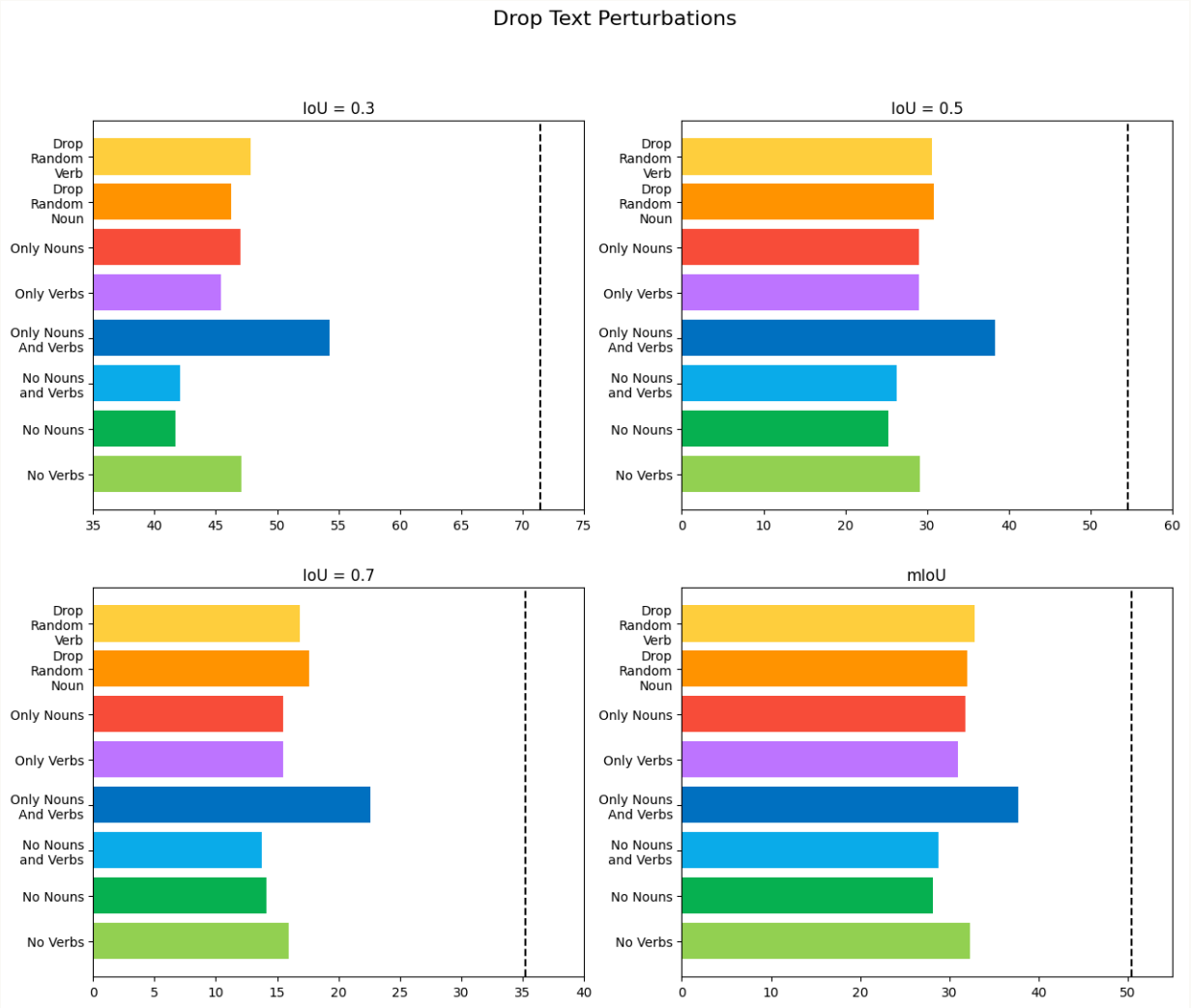
- Significant decline in performance is evident across all three subsets
- Punctuation has the worst performance

Set	Original Query	Perturbed Query
OCR	*The person sits in the chair momentarily.*	*The per8on sits in the chair momentarily.*
Prefix Swap	*Person runs to the window.*	*Person runs to the sow.*
Punctation	*Person turn a light on.*	*”””” Person turn a light on. “““*



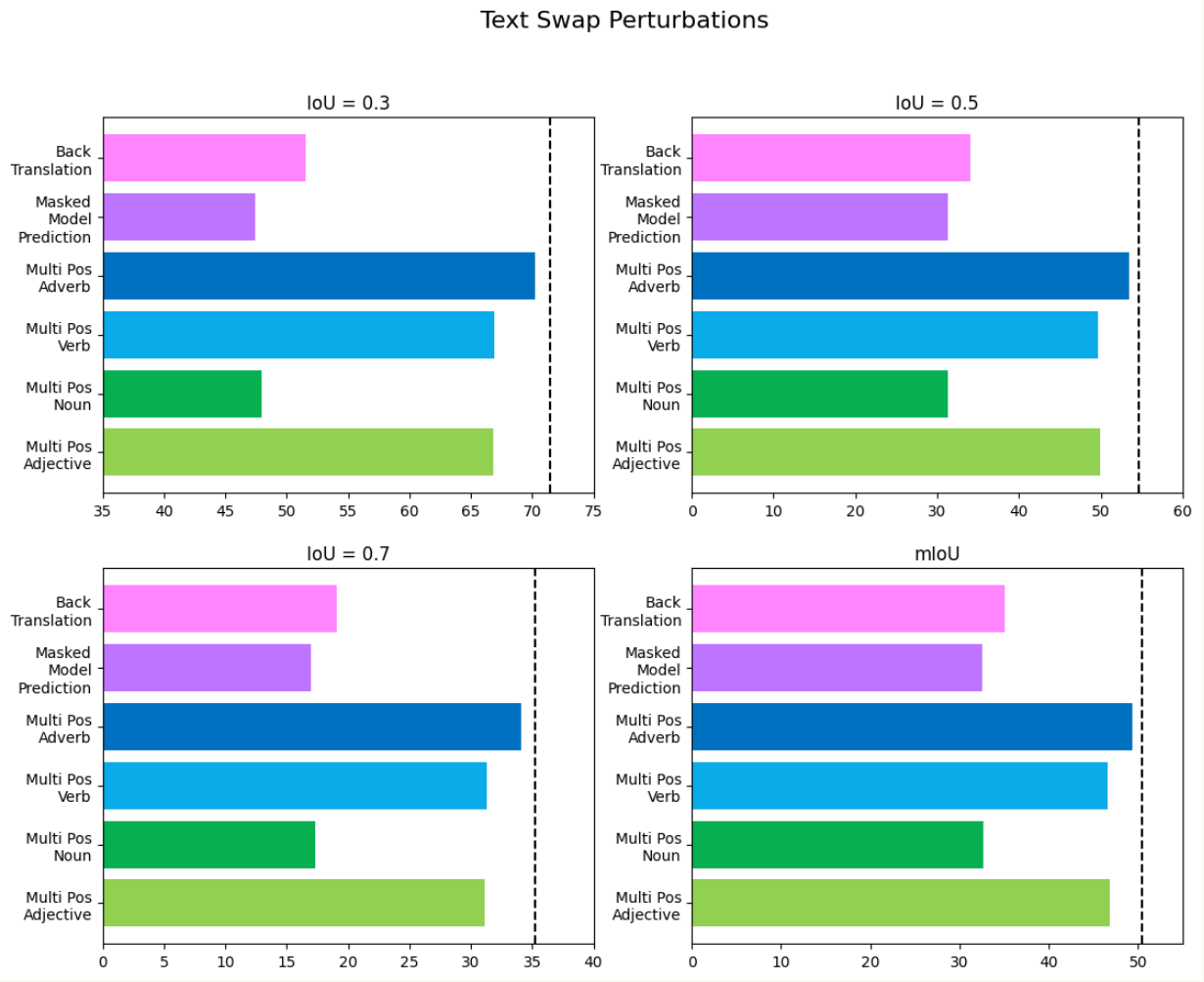
Drop Text Perturbations

Set	Original Query	Perturbed Query
Drop Rand Verb	Person turn a light on.	<UNK> turn a light on.
Drop Rand Noun	Person turn a light on.	Person <UNK> a light on.
Only Nouns	Person turn a light on.	Person <UNK> <UNK> light <UNK> <UNK>
Only Verbs	Person turn a light on.	<UNK> turn <UNK> <UNK> <UNK> <UNK>
Only Nouns & Verbs	Person turn a light on.	Person turn <UNK> light <UNK> <UNK>
No Nouns & Verbs	Person turn a light on.	<UNK> <UNK> a <UNK> on.
No Nouns	Person turn a light on.	<UNK> turn a <UNK> on.
No Verbs	Person turn a light on.	Person <UNK> a light on.



Text Swap Perturbations

Set	Original Query	Perturbed Query
Back Translation	A person is putting a picture onto the wall.	A person hangs the picture on the wall.
Masked Language Model Pred	Person turn a light on.	Person remaining a light on.
Adverb Swap	The person sits in the chair momentarily.	The person sits in the chair okay.
Verb Swap	Person takes a cup off a desk.	Person removes a cup off a desk.
Noun Swap	A person stands in the bathroom holding a glass.	A flit stands in the roomy holding a glass.
Adjective Swap	The person takes a bag from the bottom cabinet.	The person takes a bag from the hand-me-down cabinet.



Conclusion & Future Work

A quick summary of our findings and their implications and Possible avenues to expand on in the future

Conclusion

- Evaluated the generalisation capabilities of NLVL models by introducing perturbations & new datasets
- Elucidated model's reliance on specific linguistic elements (nouns & verbs)
- Highlighted the model's strengths and limitations in query processing
- Created the Charades-Ego-STA dataset to evaluate a NLVL Model’s capability to generalise on first person videos.
- Reduced the temporal distributional bias in Charades-STA dataset by creating the Charades-STA-Merged dataset, evidenced by poor performance of our bias-based models
- Provides insights into the performance and interpretability of NLVL models
- Contribute to developing more robust and fair AI systems

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Future Work

- Expand to Dependency Tag Perturbations in Query
- Explore Perturbations in Video
- Experiment with Charades-STA-Merged on different NLVL Models
- Transfer Learning for NLVL on First Person Video
- Refine Bias Mitigation Techniques

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

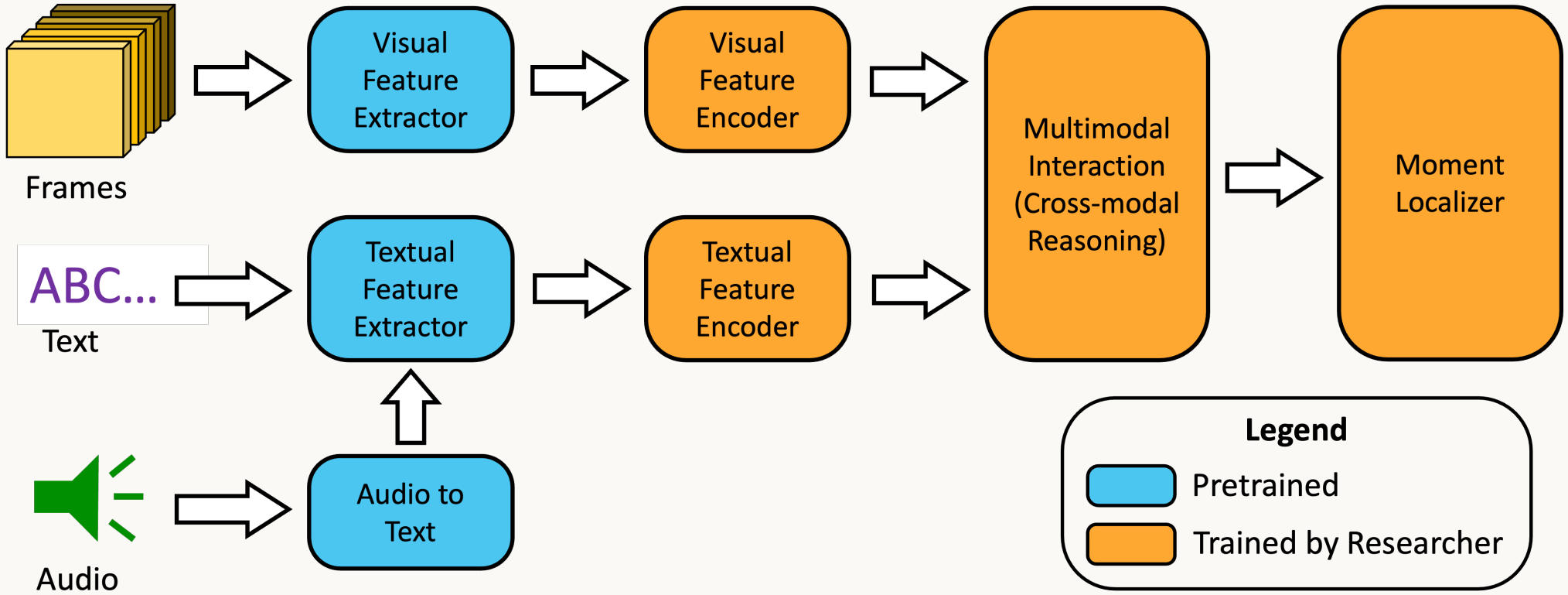
Thank You !

Please feel free to ask any questions

Available:

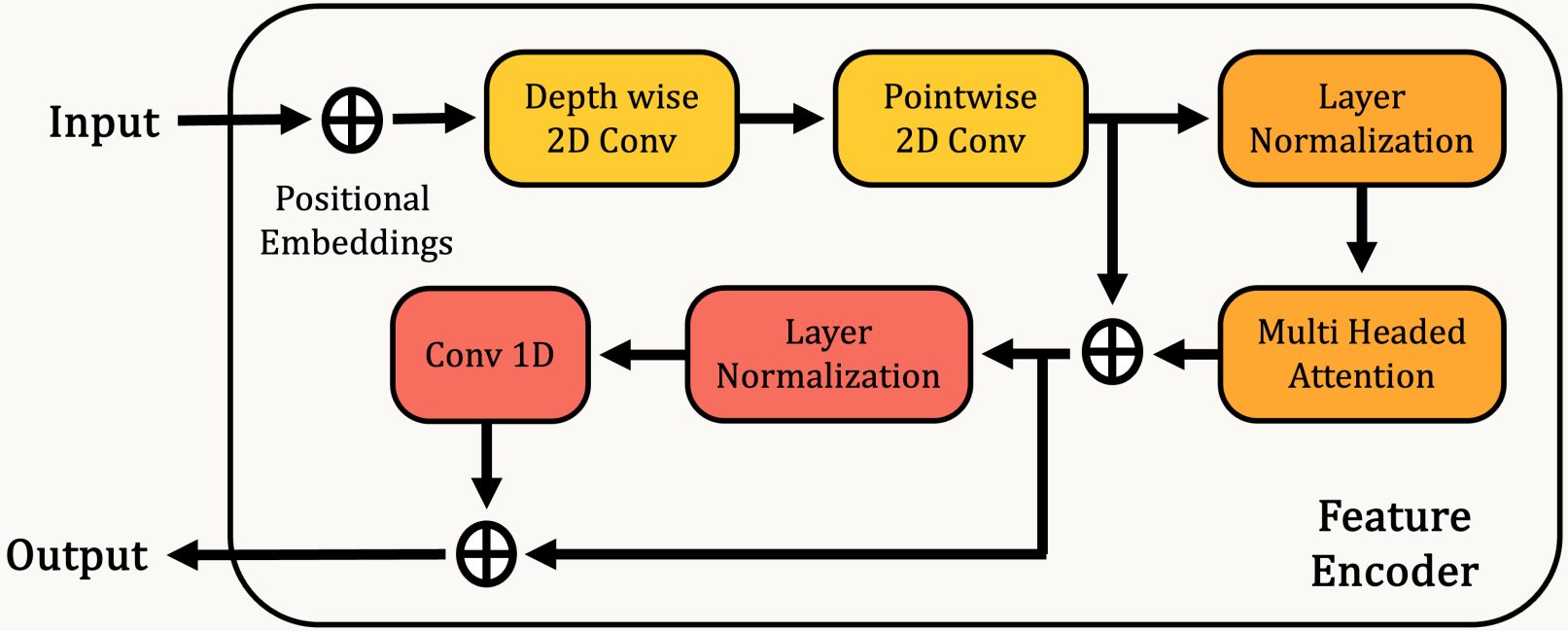
Charades-STA-Merged: Github
Charades-Ego-STA: Github
Project Report: DR-NTU

Appendix: Typical NLVL Pipeline



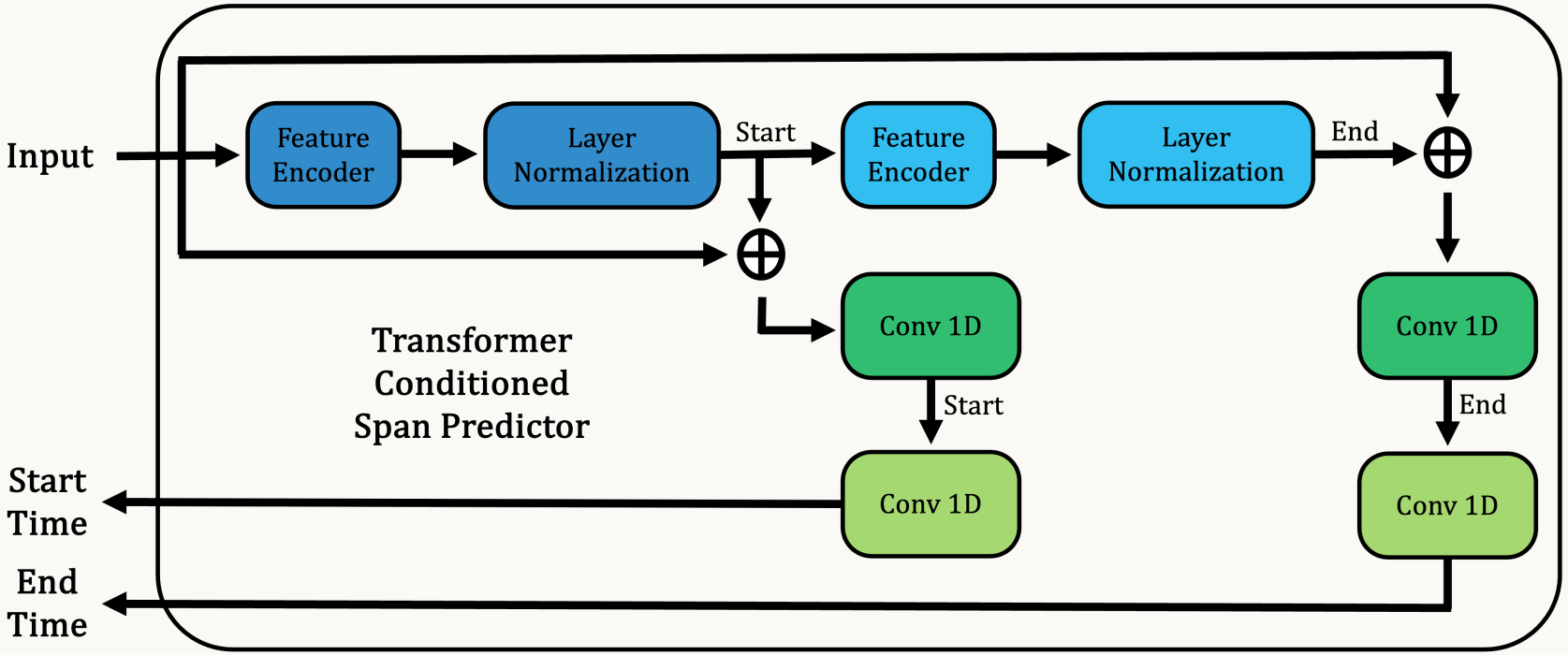
H. Zhang, A. Sun, W. Jing and J. T. Zhou, "Temporal Sentence Grounding in Videos: A Survey and Future Directions," 13 March 2023.

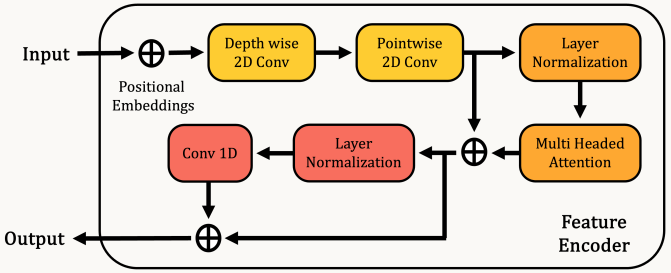
Appendix: VSLNet Feature Encoder



H. Zhang, A. Sun, W. Jing and J. T. Zhou, "Span-based Localizing Network for Natural Language Video Localization," July 2020.

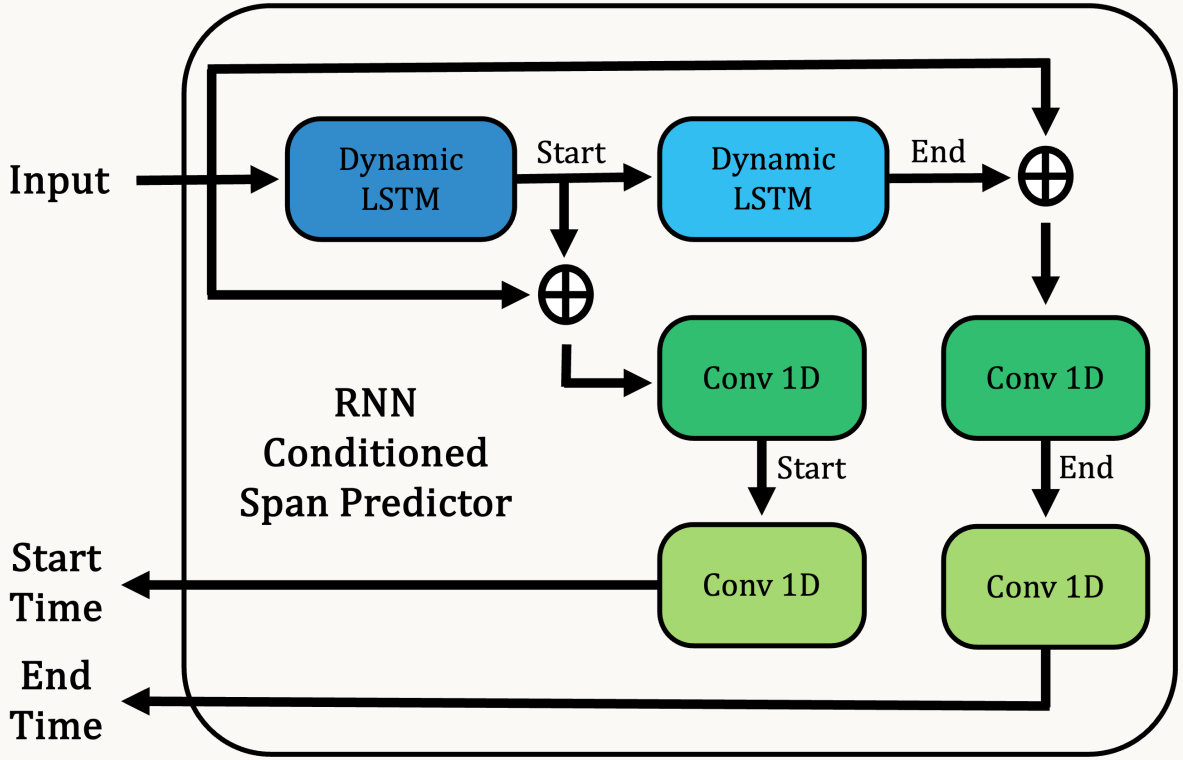
Appendix: VSLNet Span Predictor (Transformer)





H. Zhang, A. Sun, W. Jing and J. T. Zhou, "Span-based Localizing Network for Natural Language Video Localization," July 2020.

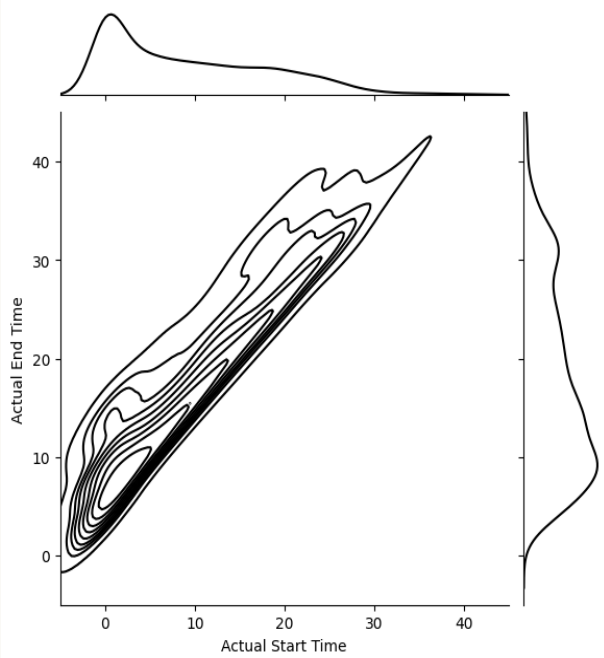
Appendix: VSLNet Span Predictor (RNN)



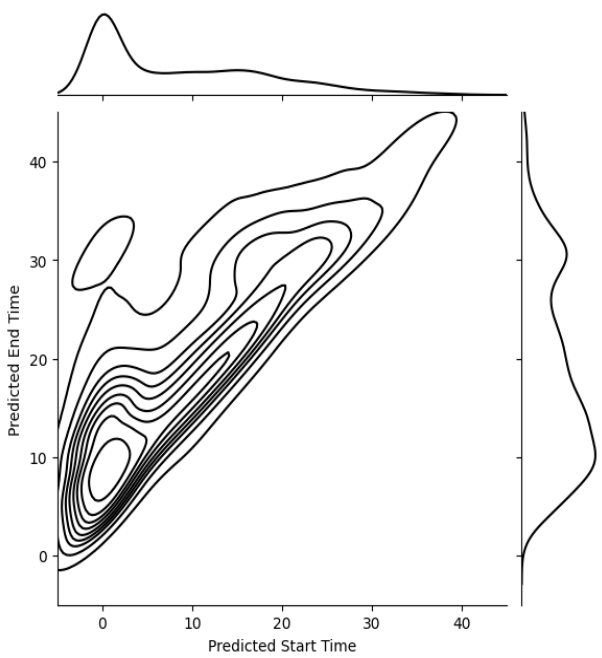
H. Zhang, A. Sun, W. Jing and J. T. Zhou, "Span-based Localizing Network for Natural Language Video Localization," July 2020.

Appendix: How models fail

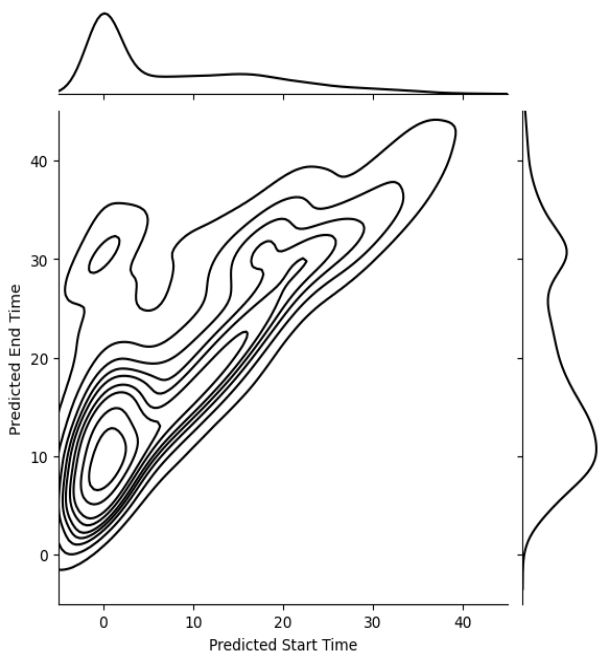
mIoU = 100



mIoU = 50.15



mIoU = 38.69



mIoU = 27.05