WHU-NERCMS @ TRECVID 2022: DEEP VIDEO UNDERSTANDING TASK

Jiahao Guo

2018302110080@whu.edu.cn

National Engineering Center for Multimedia Software
School of Computer Science, Wuhan University

December 9, 2022

Outline

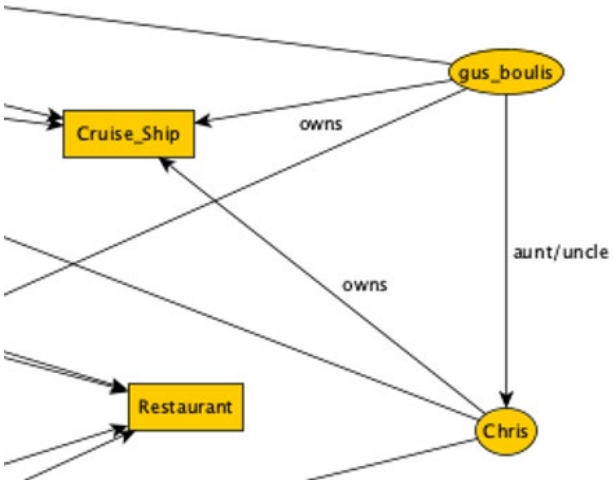
- Introduction
- Approach
- Results
- Conclusion

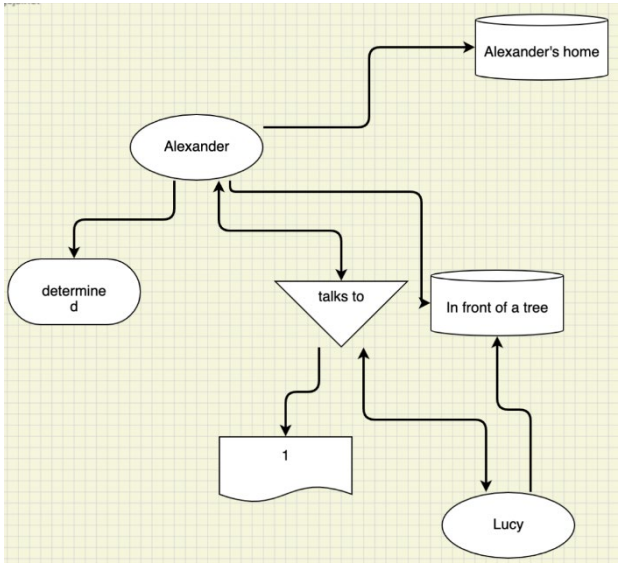
Outline

- Introduction
- Approach
- Results
- Conclusion

Introduction

- Deep Video Understanding(DVU)
  - Movie KG, entities pic, scene seg, scene KG, scene sum, vocab
  - 2 movie-level questions & 2 scene-level questions



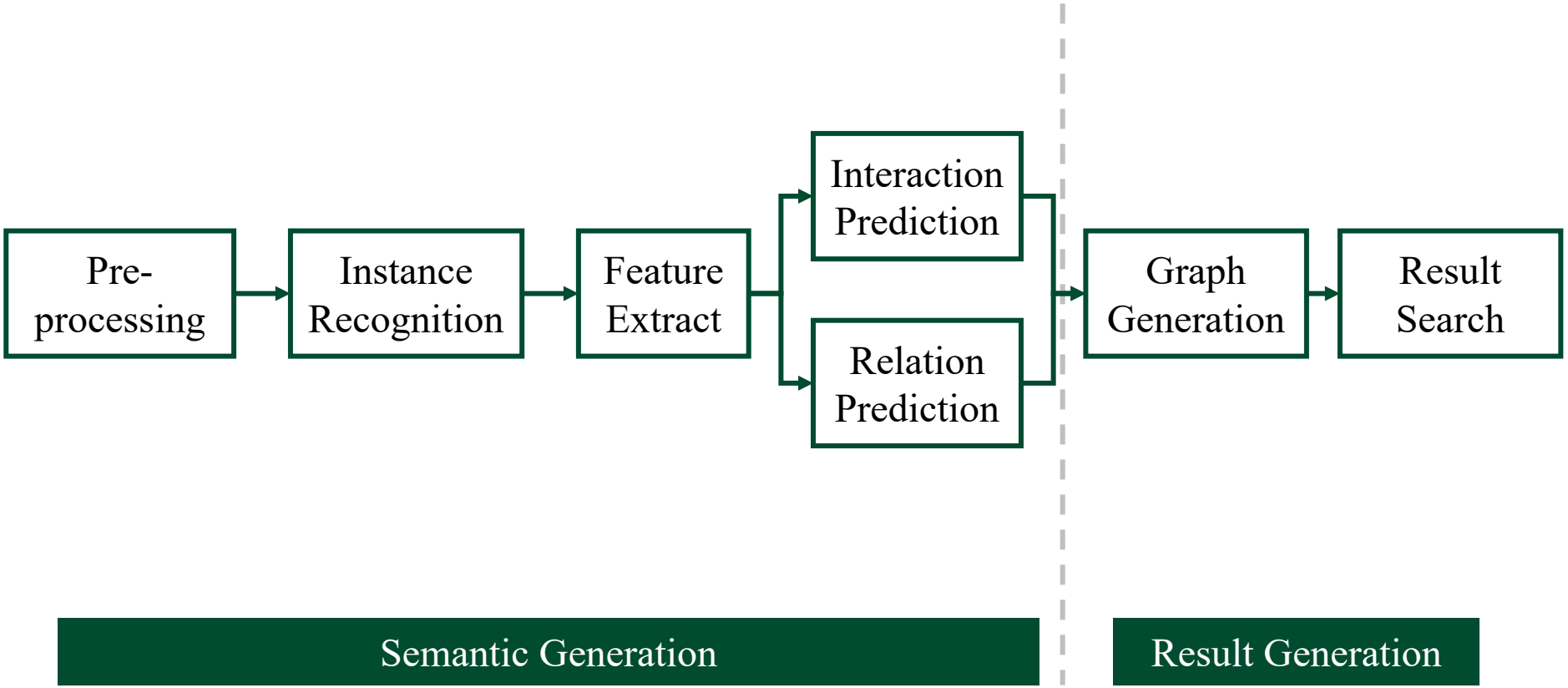


Outline

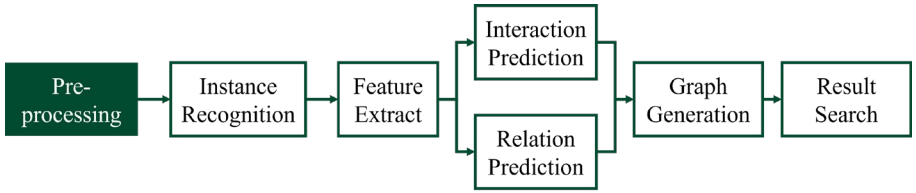
- Introduction
- Approach
- Results
- Conclusions

Approach

- Framework

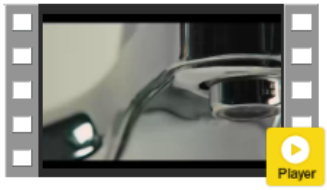


Approach



Step 1: Pre-processing

- Rename
  - Rename files of the dataset for each movie
- Segmentation
  - Scene segmentation
    a. Download scene files.
    b. Seg with timestamps locally.
  - Clip segmentation
    a. Use YouTube ASR to generate subtitles.
    b. Seg with timestamps of subtitles.

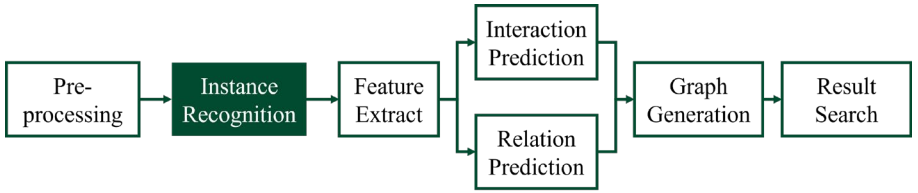


Bagman-1.webm

Bagman-2.webm

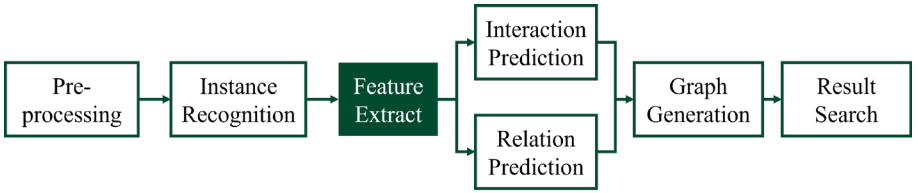
```
1
00:00:32,156 --> 00:00:33,700
- You know,

2
00:00:33,700 --> 00:00:35,326
I do a shit load of reading

3
00:00:35,326 --> 00:00:37,245
and studying and praying,
and I've come to a few
```

Approach



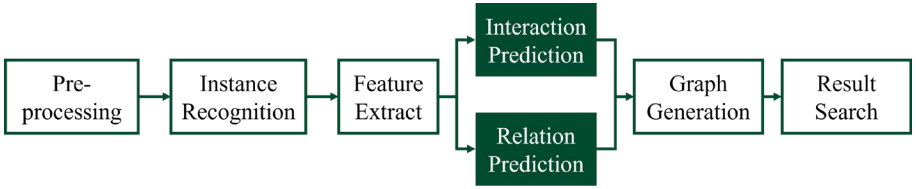
- Step 2: Instance Recognition
  - Person Recognition and Track
    - Person recognition: SCRFD + Arcface + extended face database
    - Person Track: faster RCNN + Deepsort
  - Location Recognition
    - Resnet

Approach

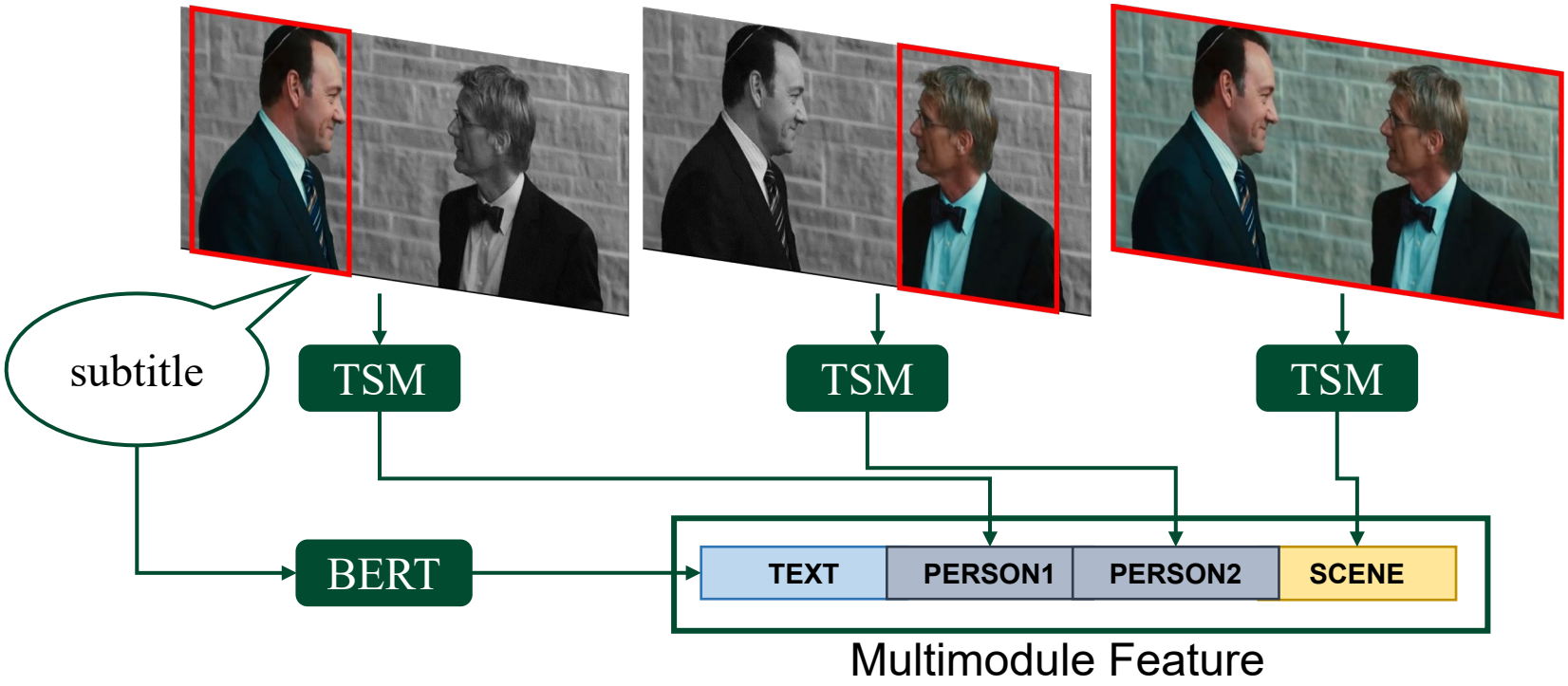


- Step 3: Feature Extract
  - Text feature
    - Bert-base extracts a feature of 768 dimensions for a clip.
  - Visual feature
    - TSM extracts a feature of 2048 dimensions for a clip.
  - Track feature
    - Unite results to generate a feature of 2048*2 dimensions for a PP/PL pair in a clip.

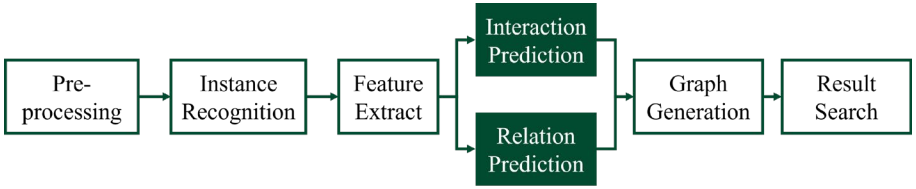
Approach



- Step 4: Interaction & Relation Prediction
  - Embedding representation

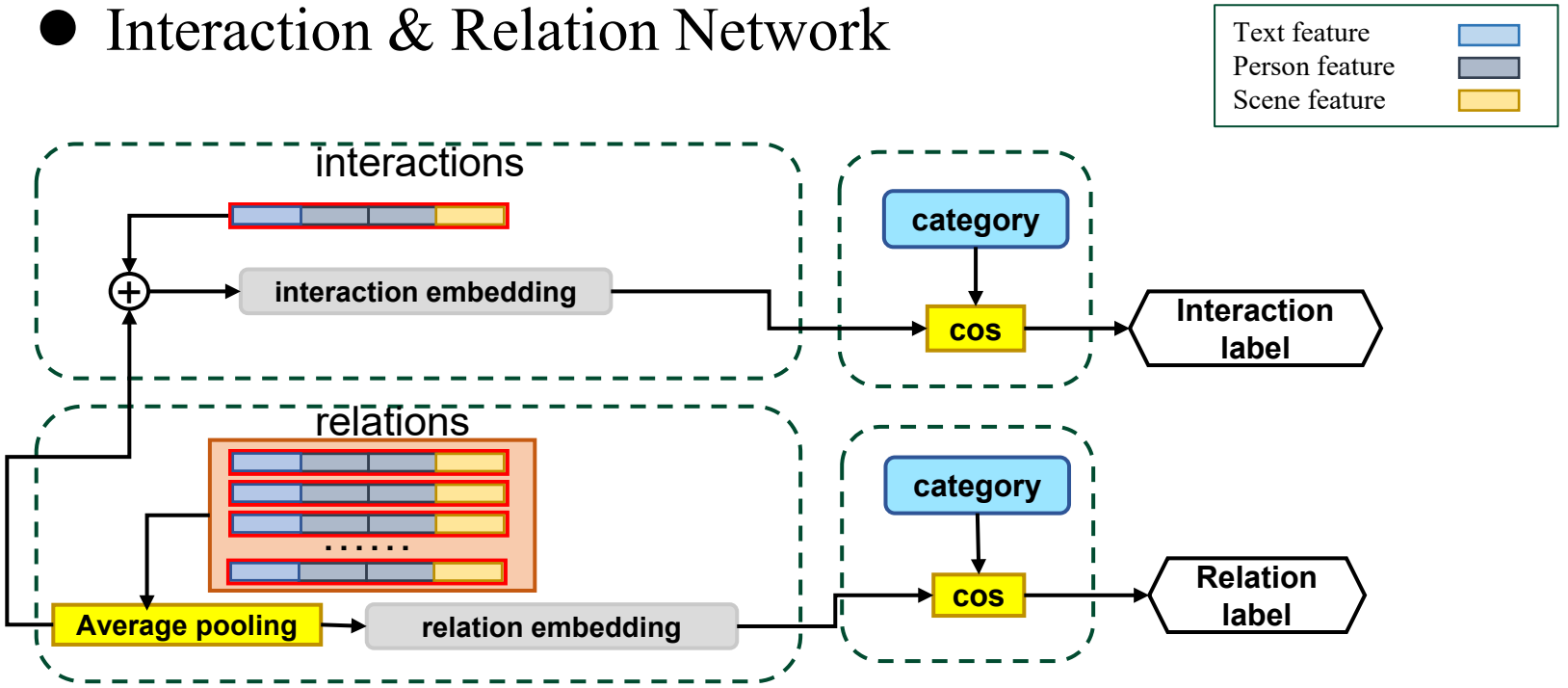


Approach

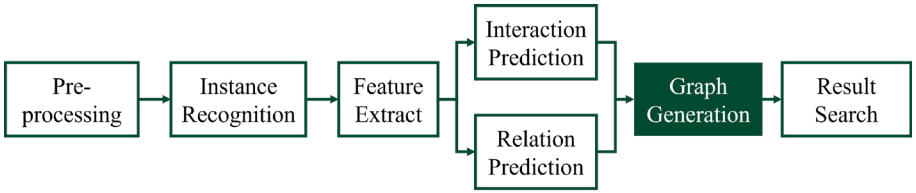


Step 4: Interaction & Relation Prediction

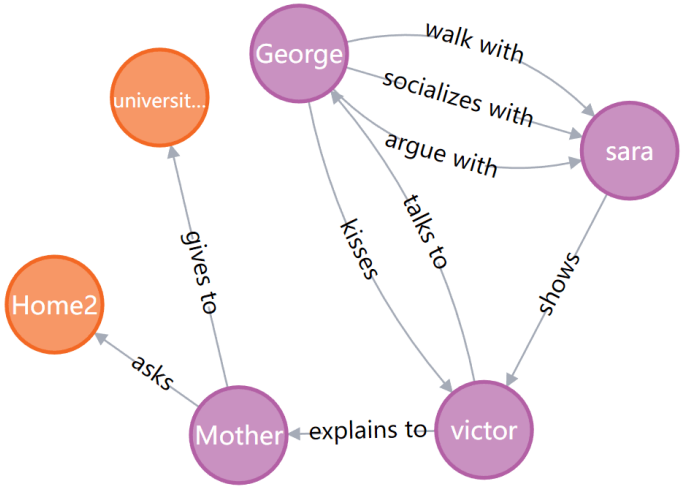
- Interaction & Relation Network

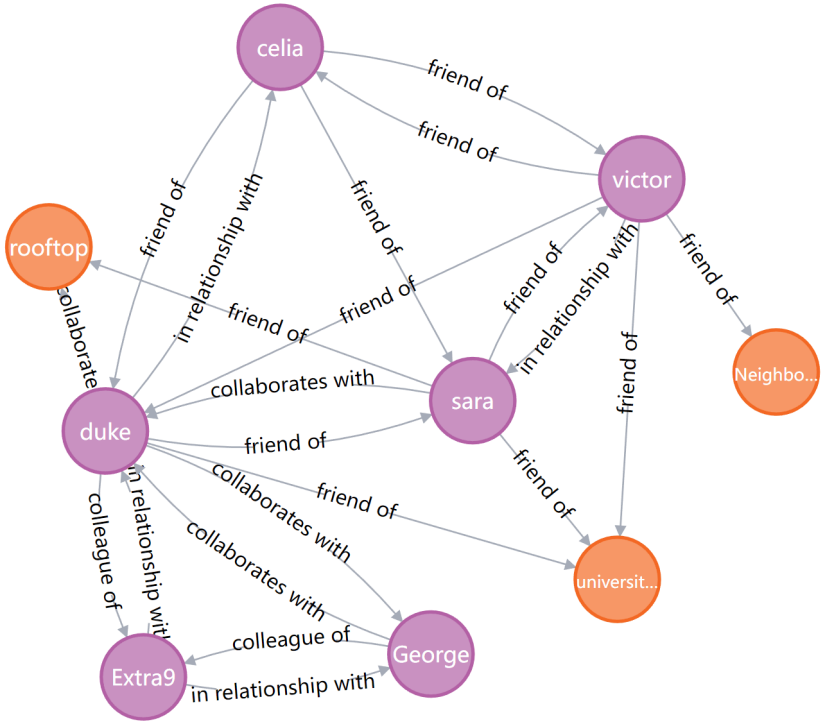


Approach

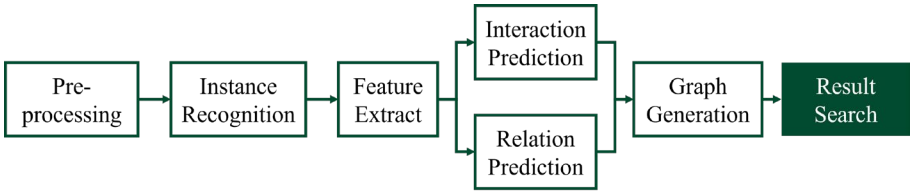


- Step 5: Graph Generation
  - neo4j





Approach



Step 6: Result Search

Movie-level Track

- Required Query Type: Question Answering (QA)

```
<DeepVideoUnderstandingTopicQuery question="1" id="1">
    <item subject="Person:Unknown_1" predicate="Relation:Lives At" object="Entity:<BLANK>"/>
    <item subject="Person:Unknown_1" predicate="Relation:Has Met" object="Person:<BLANK>"/>
    <item description="Which Person has the following Relations: Parent Of Person:Josh, Child Of Person:Solomon, In Relationship With Person:<BLANK>
```

```
MATCH (a:person_id)-[r:`works at`]->(b:location_id) where b.name='Diner'
MATCH (c:person_id)-[:`in relationship with`]->(d:person_id) where c.name in a.name
MATCH (e:person_id)-[:`lives at`]->(f:location_id) where e.name in c.name
```

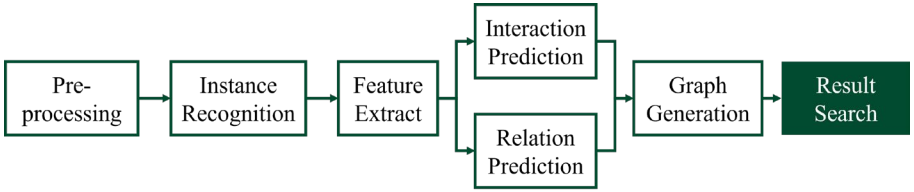
```
MATCH (e:person_id)-[r:`works at`]->(b:location_id) where b.name='Diner'
MATCH (e:person_id)-[t:`in relationship with`]->(d:person_id)
MATCH (e:person_id)-[y:`lives at`]->(f:location_id)
return toInteger(r.score)+toInteger(t.score)+toInteger(y.score), e.name
order by toInteger(r.score)+toInteger(t.score)+toInteger(y.score) DESC
```

- Optional Query Type: fill in the graph space

```
<DeepVideoUnderstandingTopicQuery question="2" id="1">
    <item subject="Person:Rabbi_Brookstein" predicate="Relation:Unknown_1" object="Person:Debbie"/>
    <item description="What is the relation / connection from Rabbi_Brookstein to Debbie?"/>
    <Answers>
        <item type="Person" answer="Apprentice Of"/>
        <item type="Person" answer="Has Met"/>
```

```
MATCH (a:person_id)-[r]->(b:person_id) where a.name='Rabbi_Brookstein' and b.name='Debbie' and r.type='rela'
    return type(r),r.score
order by toInteger(r.score)DESC
```

Approach



Step 6: Result Search

Scene-level Track

- Required Query Type: find next or previous interaction

```
<DeepVideoUnderstandingTopicQuery question="2" id="1">
    <item subject="Person:Debbie" scene="18" predicate="Interaction:talks to" object="Person:Co-Worker"/>
    <item description="In Scene 18, Debbie talks to Co-Worker. What is the  immediate next / following interation between Co-Worker and Debbie, in  scene 18?"/>
    <Answers>
        <item type="Interaction" scene="18" answer="greets"/>
        <item type="Interaction" scene="18" answer="hits"/>
```

```
match(a:person_id)-[r]->(b:person_id) where r.scence='18' and a.name='Debbie' and b.name='Co-Worker'
return id(r),type(r) order by id(r)
```

- Optional Query Type: find the unique scene

```
<DeepVideoUnderstandingTopicQuery question="1" id="1">
    <item subject="Scene:Unknown_1" predicate="Interaction:asks" object="Person:BLANK"/>
    <item subject="Scene:Unknown 1" predicate="Interaction:shows" object="Person:BLANK"/>
    <item description="Which Unique Scene contains the following Interactions: asks, shows, explains to, shoots, shoots"/>
```

```
match()-[r:`asks`]->()
match()-[t:`shows`]->()where t.scence=r.scence
match()-[y:`explains to`]->() where y.scence=t.scence

return toInteger(y.score)+toInteger(t.score),t.scence
order by  toInteger(y.score)+toInteger(t.score) DESC
```

Outline

- Introduction
- Approach
- Results
- Conclusion

Results

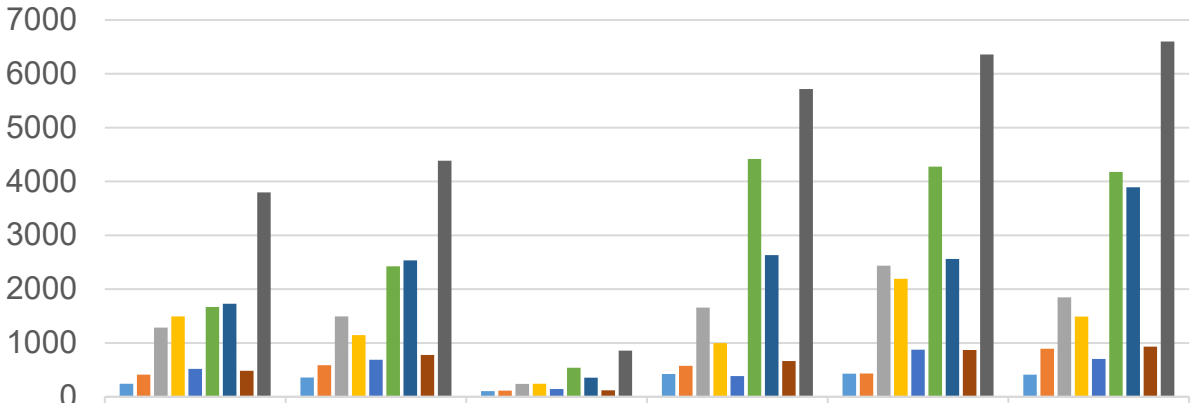
- Overall Result
  - Movie-level & Scene-level

Movie-level	Result	Scene-level	Result
run_1	28.9	run_1	11.1
run_2	9.6	run_1	3.1

  - Discussion
    - Run_2: pretrained LIREC model
    - Run_1: trained model on self-labeled data

Results

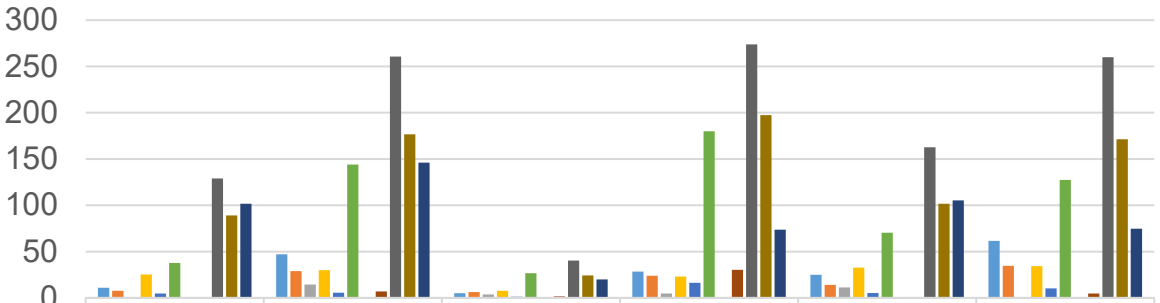
Interaction



	losing_ground	little_rock	like_me	liberty_kid	chained_for_life	calloused_hands
ask	244	358	105	423	430	412
shows	410	586	113	576	431	893
explains to	1286	1493	239	1658	2434	1847
talks to	1494	1146	242	996	2191	1491
gives to	519	689	143	386	876	704
argue with	1669	2422	541	4418	4274	4176
walk with	1728	2534	354	2633	2560	3892
kisses	482	778	120	664	871	932
socializes with	3797	4385	857	5716	6359	6598

Results

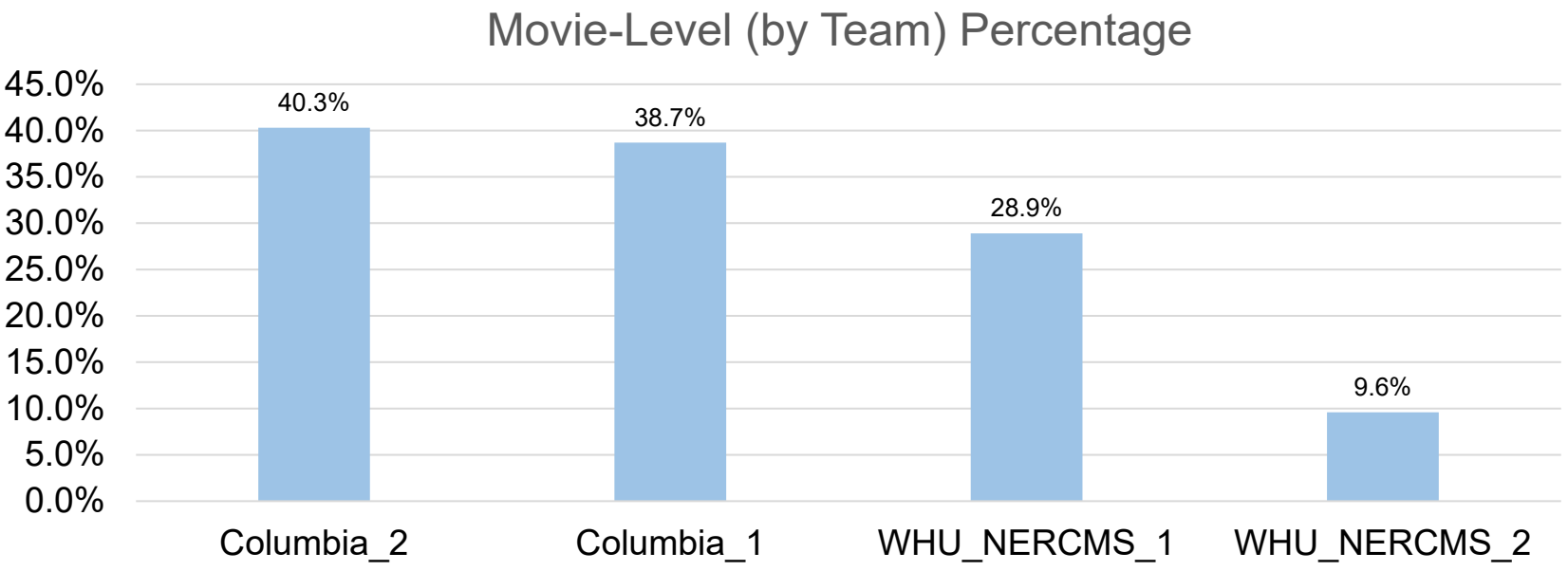
Relation



	losing_ground	little_rock	like_me	liberty_kid	chained_for_life	calloused_hands
has met	11	47	5	28	25	62
socialises at	8	29	6	24	14	35
siblings of	0	14	4	5	11	0
works at	25	30	8	23	33	34
child	5	6	1	16	5	10
in relationship with	38	144	27	180	70	127
lives at	0	0	0	0	0	0
residence of	1	7	2	30	1	5
friend of	129	261	40	274	163	260
colleague of	89	177	24	197	102	171
collaborates with	102	146	20	74	105	75

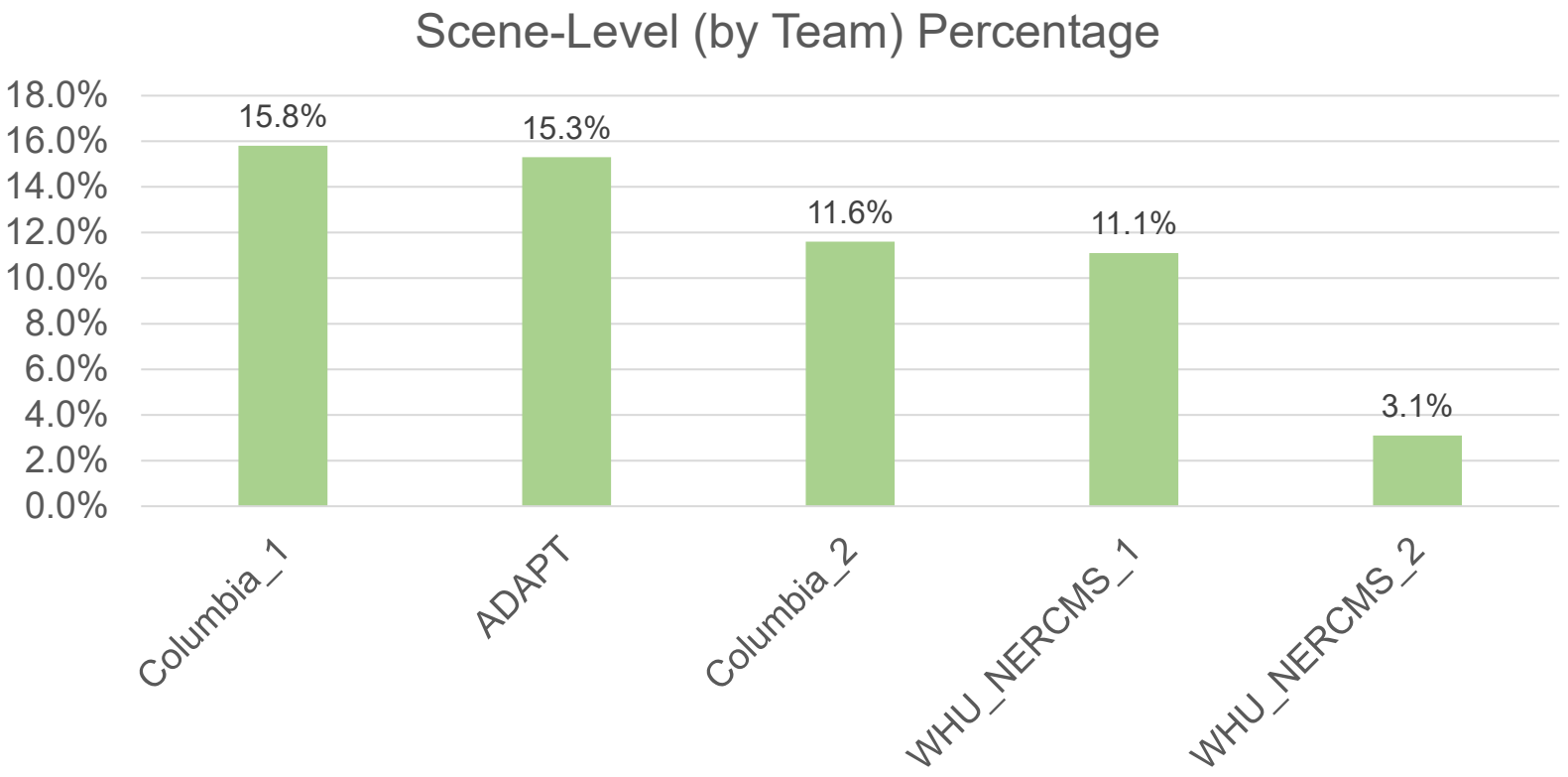
Results

- Movie-level Result



Results

- Scene-level Result



Outline

- Introduction
- Approach
- Results
- Conclusions

Conclusion

- Some action may occur without text
  - use MulT (multimodal transformer) for concatenate.
- Annotation is inadequate
  - Pretrain for augmentation: add data with high degree of confidence to train set (cycle).

Thanks for your time!

Hubei Key Laboratory of Multimedia and Network Communication Engineering
National Engineering Center for Multimedia Software
School of Computer Science, Wuhan University

(reported on December 9, 2022)