TRECVID 2022 DEEP VIDEO UNDERSTANDING INTRODUCTION AND TASK OVERVIEW

Keith Curtis
National Institute of Standards and Technology

George Awad
National Institute of Standards and Technology

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST, or the U.S. Government.

Table of Contents

Task Goals

- Analyze long duration videos holistically.
- Exploit all available modalities (audio, video, image, & text) to analyze both visual and non-visual elements.
- As the movies domain data can simulate the real world, many lessons learned are expected to benefit different kinds of applications

Task Definition

- Given:
  - Whole original **movie** (e.g 1.5 - 2hrs long)
  - **Image snapshots** of main entities (persons, locations, and concepts) per movie
  - **Ontology** of relationships, interactions, locations, and sentiments used to annotate each movie at global movie-level (relationships between entities) as well as on fine-grained scene-level (scene sentiment, interactions between characters, and locations of scenes)

- Generate a knowledge-base of the main actors and their relations (such as family, work, social, etc.) over the whole movie, and of interactions between them over the scene level.

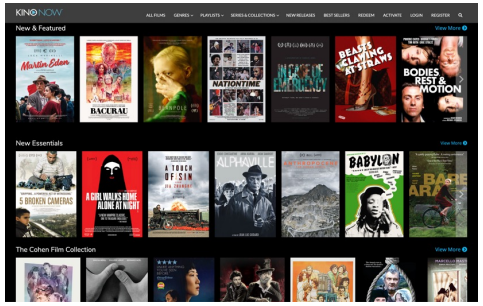
- The task supported two auto generated query types on the **movie-level** as well as on **scene-level** per movie.

Data

- Training Set comprised of 14 Creative Commons (CC) licensed movies.
  - Long duration videos with a self-contained storyline.
  - Videos range from 18 minutes in length to 109 minutes.
- Test Set comprised of 6 movies licensed from Kinolorber*.
  - Long duration videos with a self-contained storyline.
  - Videos range from 79 minutes in length to 92 minutes.

INTERNET ARCHIVE

vimeo



*https://kinolorberedu.com/

NIST

Annotation Framework

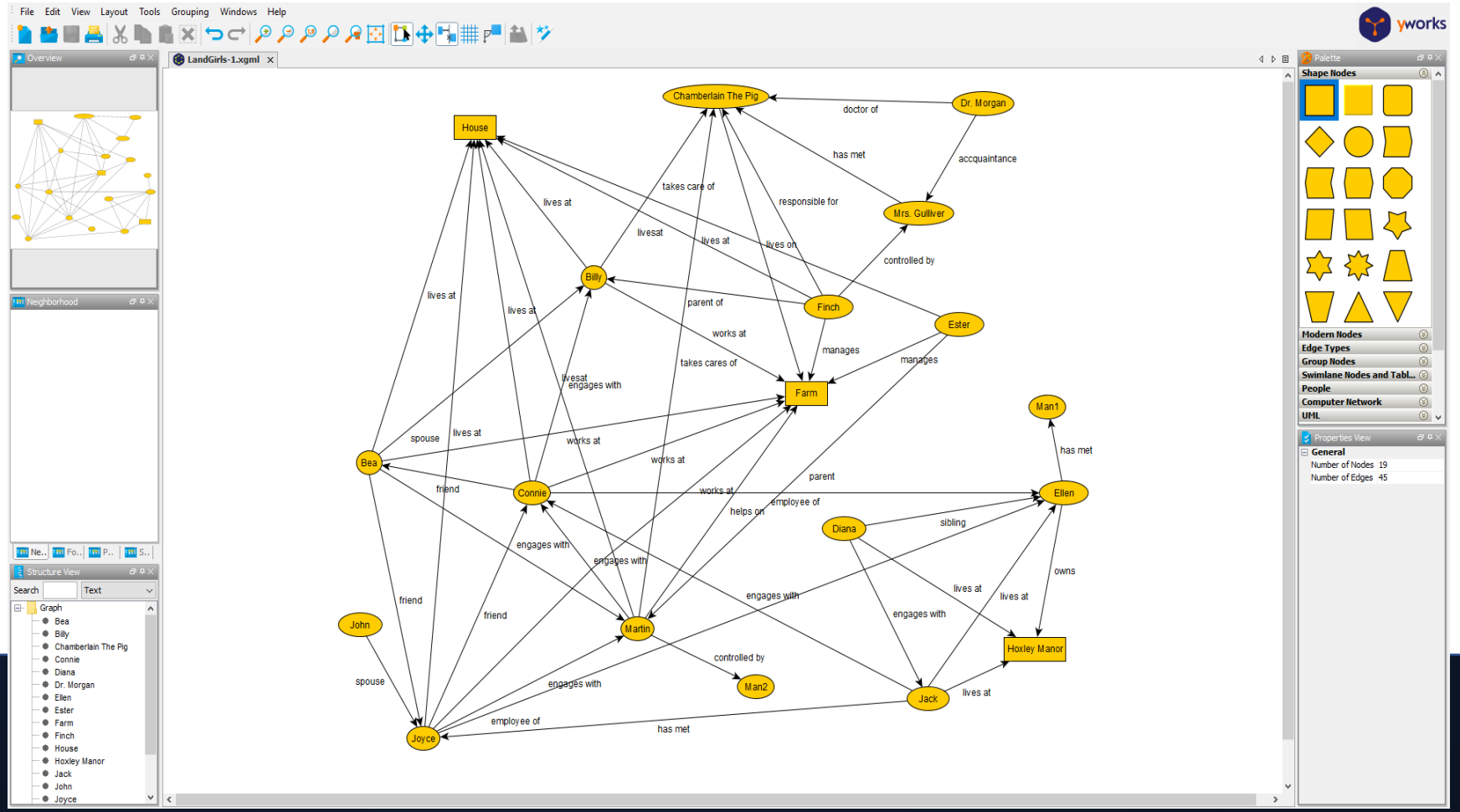
- Movies are first divided into scenes.
- A set of dedicated annotators were hired to work with us on the annotation framework[1].
- Annotators watch full movies, isolate and take images of main characters, places, & concepts. Draw Knowledge Graph (KG) of full movie using yEd* graphing tool.
- Annotators watch individual scenes, and draw KG over the scene level recording interactions between characters, chronological order of such, scene sentiments, relationships, character's emotional states, and a natural language description.

[1] Loc, E., Curtis, K., Awad, G., Rajput, S., & Soboroff, I. (2022). Development of a MultiModal Annotation Framework and Dataset for Deep Video Understanding. *P-VLAM*, 12.

* https://www.yworks.com/products/yed

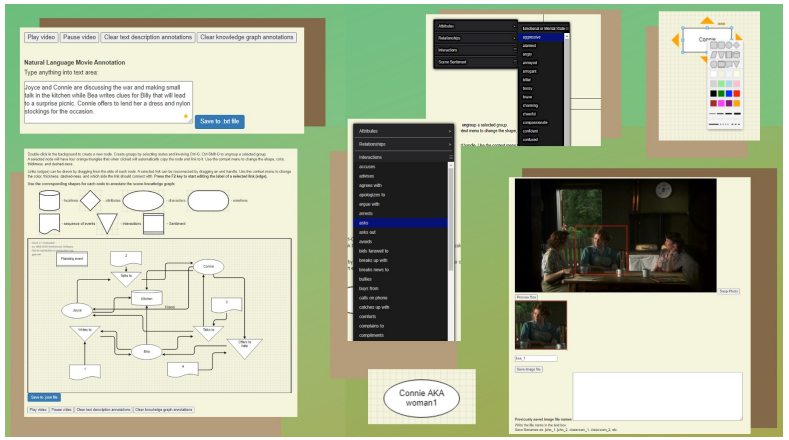
Annotation: Movie-level

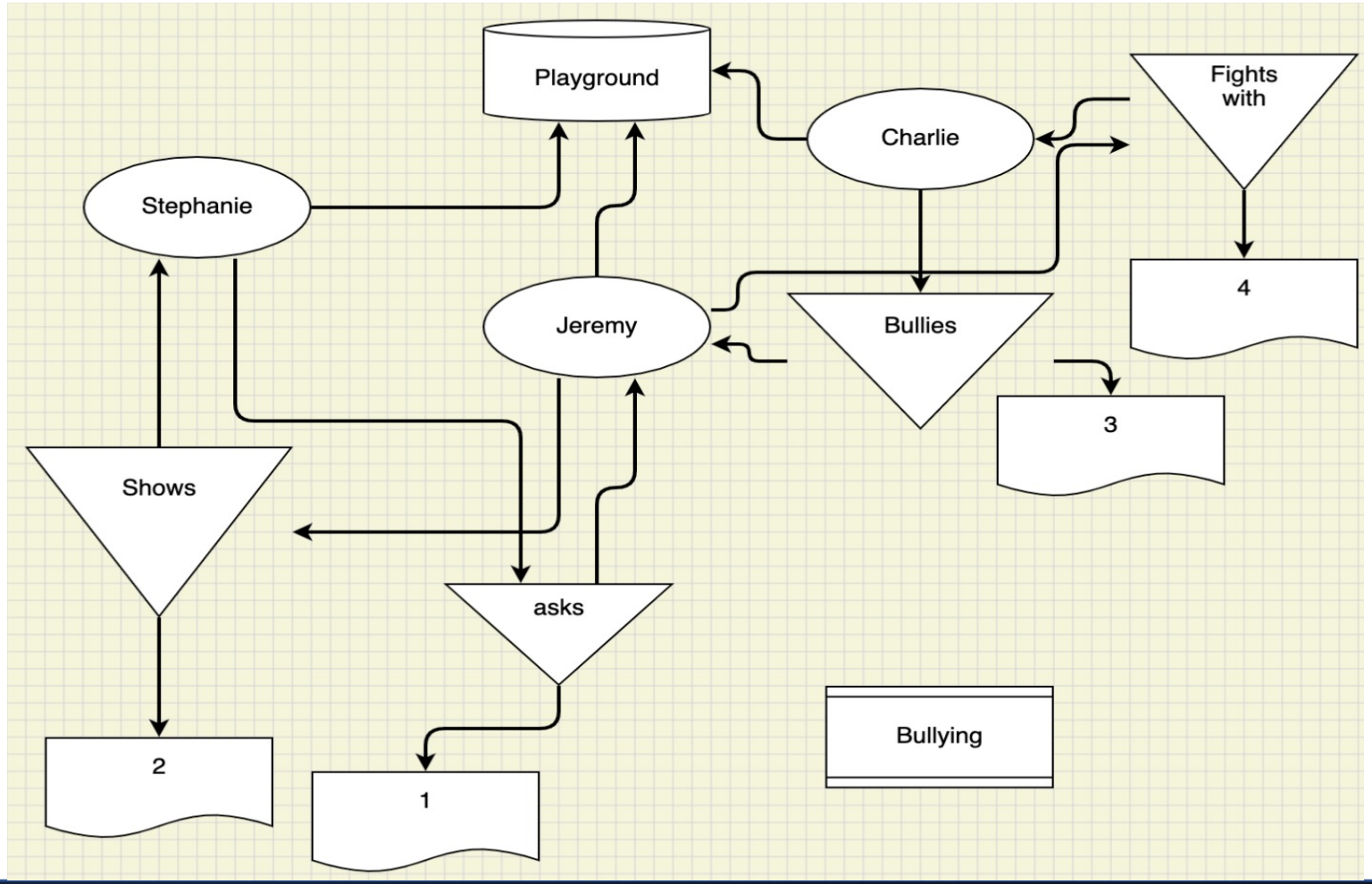
- KG annotates relations between main characters, locations, and entities.
- XGML graph file is processed later for query generation



Annotation: Scene-level

- KG annotates interactions, attributes, and relations between characters.
- Natural language text descriptions are also provided for each scene.





Queries: Movie-level

1. Question Answering (**Required**): This query type represents questions on the resulting KG in the form of multiple-choice questions.

2. Fill in the graph space (Optional): Given a list of people, entities, and/or relationships for certain nodes, where some nodes are replaced by variables X, Y, etc., solve for X, Y etc.

Queries: Scene-level

1. Find the next or previous interaction (**Required**): Given a scene number *a*, and an interaction *i* between two characters *x* & *y*, what is the immediate next or previous interaction, in scene *b,* between *x* and *y*?

2. Find the unique scene (Optional): Given a full, inclusive list of interactions, unique to a specific scene in the movie, teams should find which scene this is.

Query Samples: Movie-level

```
▼<DeepVideoUnderstandingTopicQuery question="3" id="2">
   <item subject="Person:Manny" predicate="Relation:Works At" object="Entity:Unknown_2"/>
   <item description="Where does Manny work?"/>
  ▼<Answers>
     <item type="Entity" answer="Private_Plane"/>
     <item type="Entity" answer="Beach_House"/>
     <item type="Entity" answer="Bathroom"/>
     <item type="Entity" answer="Gym"/>
     <item type="Entity" answer="City"/>
     <item type="Entity" answer="office_building"/>
   </Answers>
 </DeepVideoUnderstandingTopicQuery>
```

Visual modality helps to answer the query






**All images are under CC license

Query Samples: Scene-level

```
▼<DeepVideoUnderstandingTopicQuery question="4" id="4">
   <item subject="Person:Jack" scene="28" predicate="Interaction:watches" object="Person:Pam"/>
   <item description="In Scene 28, Jack watches Pam. What is the immediate prior / previous interation between Jack and Pam, in scene 19?"/>
  ▼<Answers>
     <item type="Interaction" scene="19" answer="shows"/>
     <item type="Interaction" scene="19" answer="asks"/>
     <item type="Interaction" scene="19" answer="reassures"/>
     <item type="Interaction" scene="19" answer="talks to"/>
     <item type="Interaction" scene="19" answer="negotiates with"/>
     <item type="Interaction" scene="19" answer="socializes with"/>
   </Answers>
 </DeepVideoUnderstandingTopicQuery>
```

Audio modality helps to answer the query

Jack

Pam

Scene 19

** Images and video clip are under CC license

Metrics

Movie-Level

- Question answering : correct answers/total questions.
- Fill in Graph questions : Mean Reciprocal Rank (MMR).

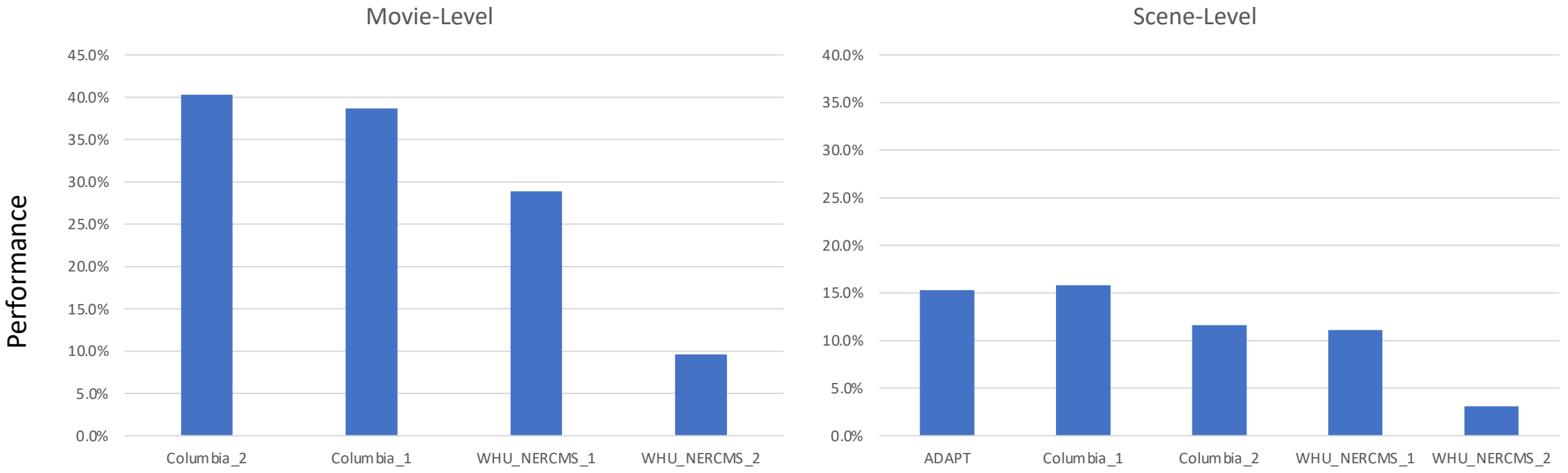
Scene-Level

- Next / Previous interaction questions : correct answers/total questions.
- Find unique scene : Mean Reciprocal Rank (MMR).

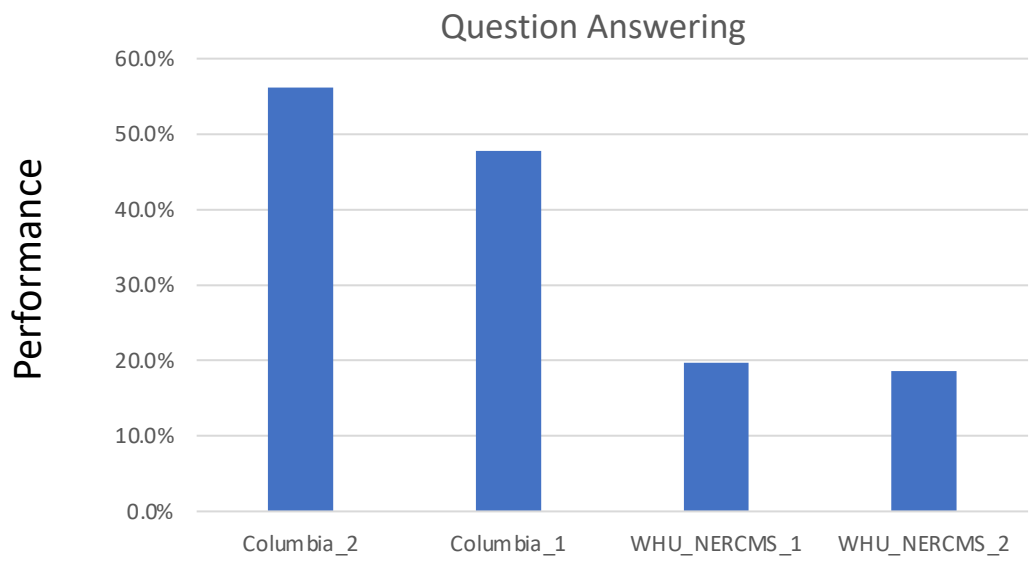
DVU 2022: 3 Finishers (out of 13)

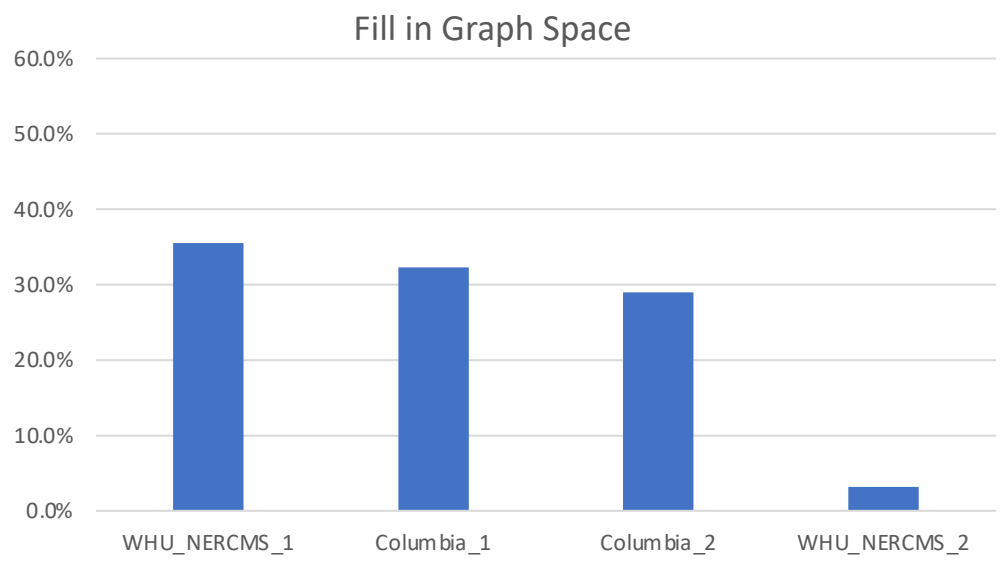
Team	Organization
Adapt	ADAPT Research Centre, Dublin City University
columbia_graphen	Graphen, Inc., Columbia University
WHU_NERCMS	National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan City, Hubei Province, China

Results Summary by run

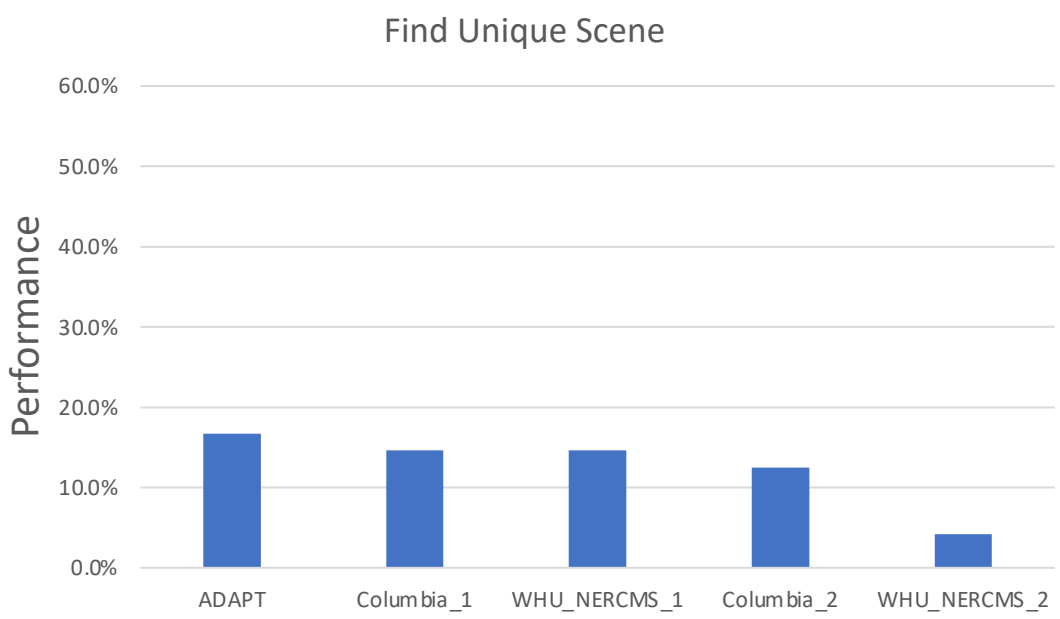


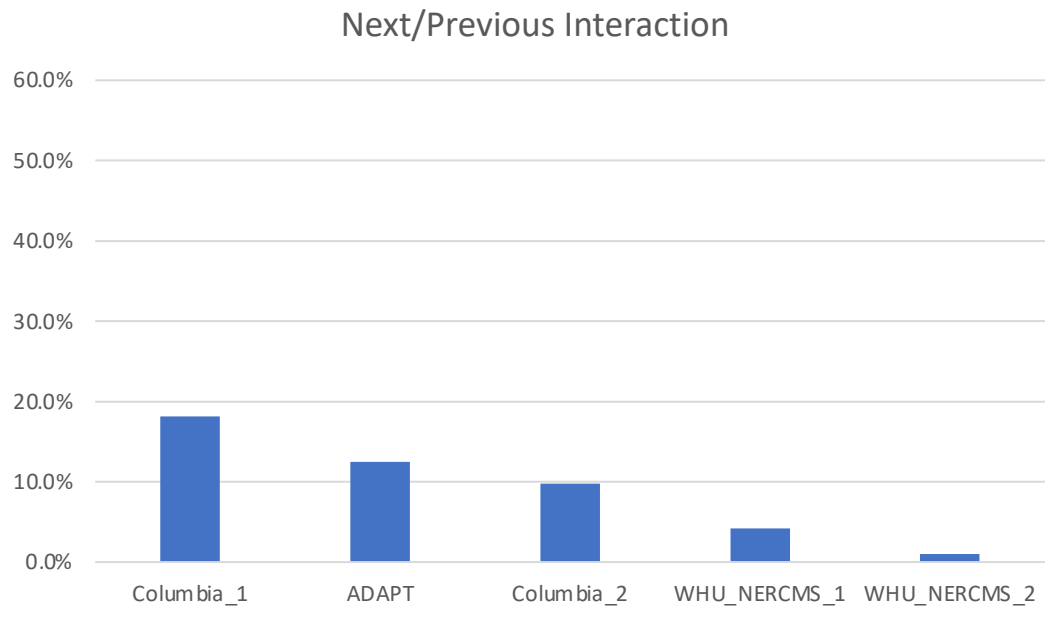
Results by query types : Movie-Level



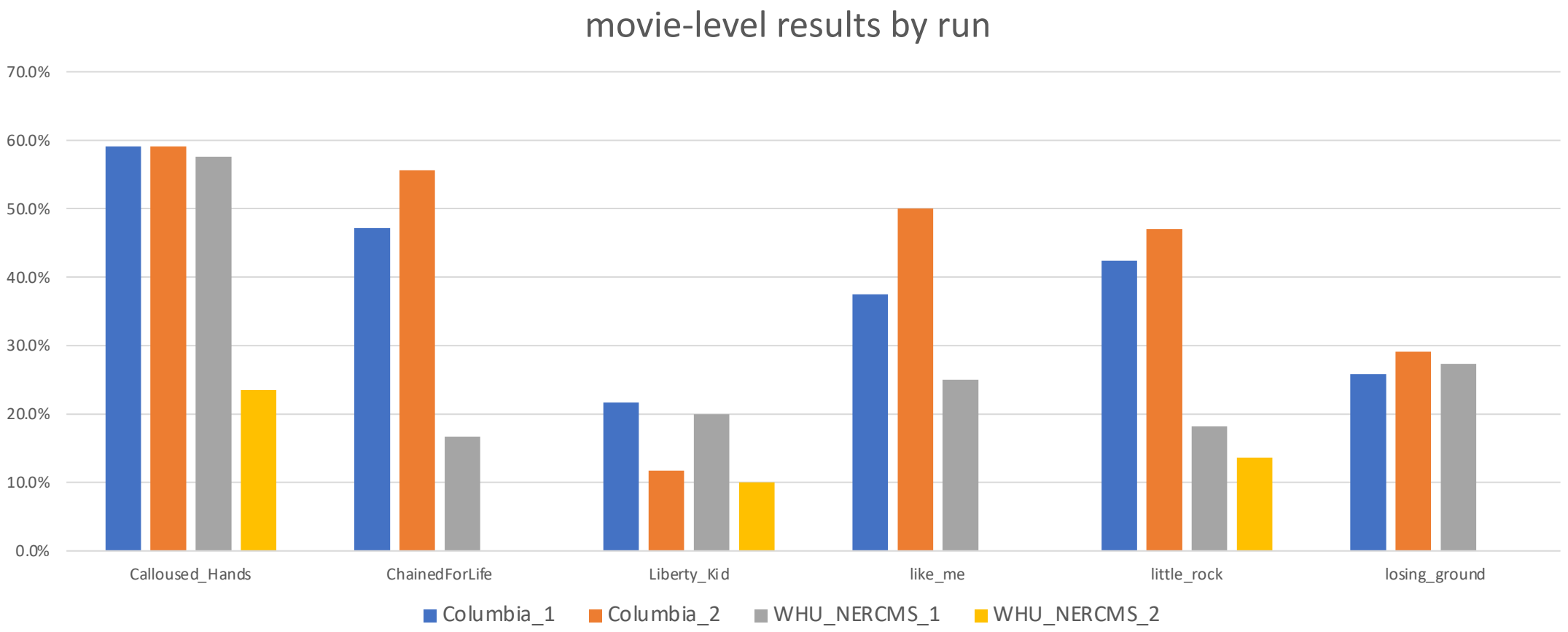


Results by query types : Scene-Level

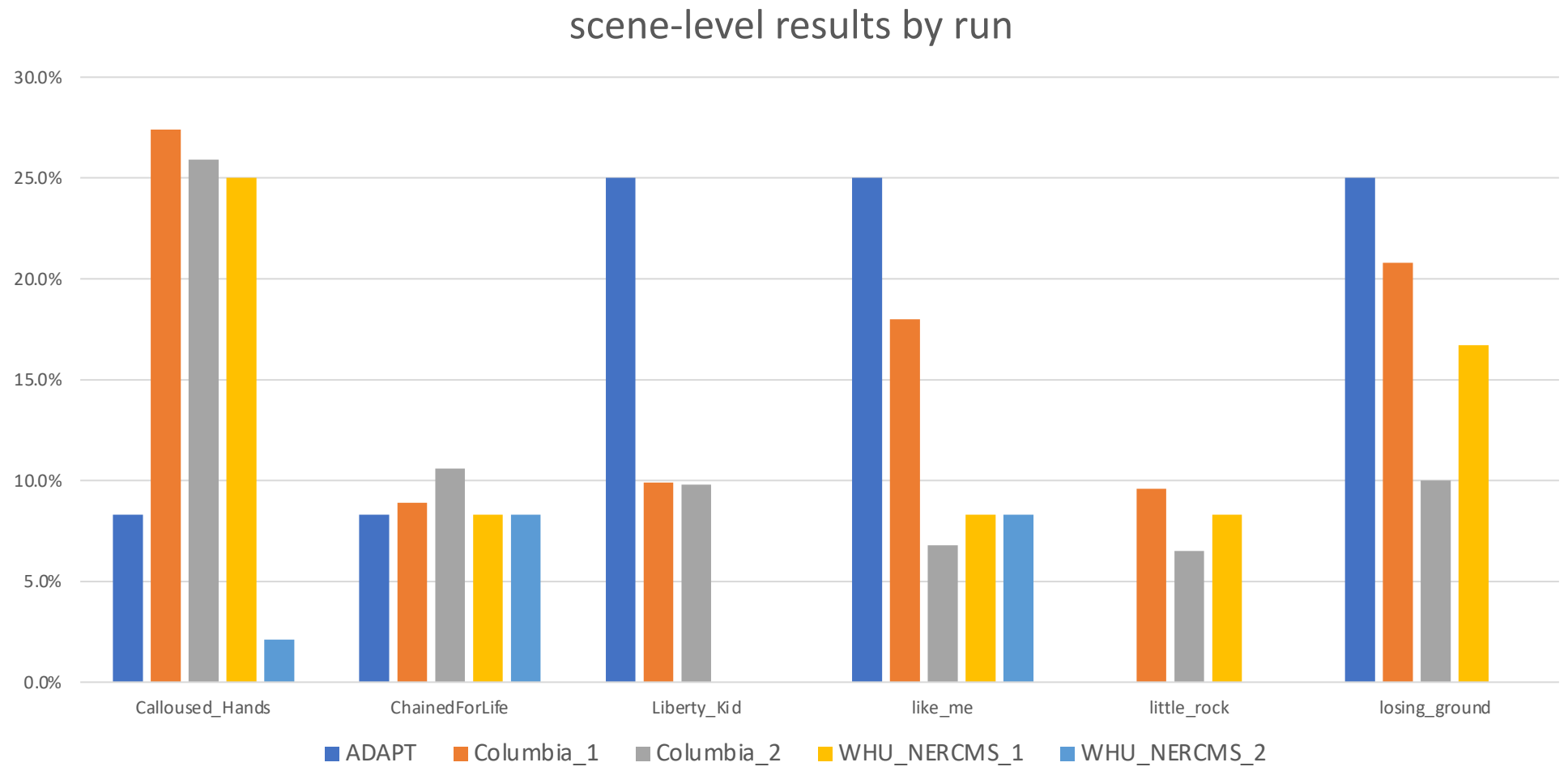




Results by runs



Results by runs



Conclusions

1. Submissions for movie-level queries scored much higher than for scene-level queries, indicating that the movie-level queries were easier or movie-level algorithms were more successful than the scene algorithms.

2. Movie-level Question Answering queries scored higher than Fill in the Graph Space queries, also indicating that these were slightly easier.

3. Scene-level Find Next or Previous Interaction queries scored close results to Find the Unique Scene queries.

4. Queries for the movie ‘Calloused Hands’ scored higher than any other movie. ‘Liberty Kid’ was among the lowest scoring.

Conclusions

5. 13 teams registered for this year's task. Out of these, 3 teams submitted runs. We would wish to see an improvement on this in subsequent years.

6. Improvements to the dataset for this task is an ongoing process, test-set movies from this year's task will be added to the training-set for next year's task, and additional new movies will be annotated and used for the test-set.