TRECVID 2023 DEEP VIDEO UNDERSTANDING

TASK OVERVIEW

Keith Curtis

Technological University of the Shannon, Ireland

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 & Definition
- Data
- Annotation Framework
- Topics (Queries)
- Participating Teams
- Evaluation and Results
- General Observation

Task Goals

- Analyze & understand long duration videos holistically.
- Exploit all available modalities (audio, video, image, & text) to comprehend 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 real-world applications

Task Definition

- Given:
  - Whole raw **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.
- 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 query types on the **movie-level** and **scene-level** per movie.

Data

- Long duration videos with a self-contained storyline.

- Training Set : 19 movies (~ 25 hrs)
  - 14 Creative Commons (CC) movies
  - 5 licensed Kinolorber movies
  - Videos range from 18 minutes in length to 109 minutes

- Test Set : 5 movies (~ 7.5 hrs) licensed from Kinolorber*
  - Videos range from 79 minutes in length to 114 minutes.

*https://kinolorberedu.com/

Data – Training Set (19 CC movies) ~ 25 hrs

Movie	Genre	Length
Honey	Romance	86 minutes
Let's bring back Sophie	Drama	50 minutes
Nuclear Family	Drama	28 minutes
Shooters	Drama	41 minutes
Spiritual Contact	Fantasy	66 minutes
Super Hero	Fantasy	18 minutes
The Adventures of Huckleberry Finn	Adventure	106 minutes
The Big Something	Comedy	101 minutes
Time Expired	Comedy / Drama	92 minutes
Valkaama	Adventure	93 minutes
Bagman	Drama / Thriller	107 minutes
Manos	Horror	73 minutes
Road To Bali	Comedy / Musical	90 minutes
The Illusionist	Adventure / Drama	109 minutes

Movie	Genre	Length
Calloused Hands	Drama	92 minutes
Chained For Life	Comedy / Drama	88 minutes
Liberty Kid	Drama	88 minutes
Like Me	Horror / Thriller	79 minutes
Losing Ground	Comedy / Drama	81 minutes

2022 testing set

Data – Test Set (5 movies licensed from KinoLorberEdu*) ~ 7.5 hrs

Movie	Genre	Length
Archipelago	Drama	114 minutes
Bonneville	Drama	93 minutes
Heart Machine	Drama	85 minutes
Little rock	Drama	82 minutes
Memphis	Drama	79 minutes

https://www.kinolorberedu.com/

Annotation Framework

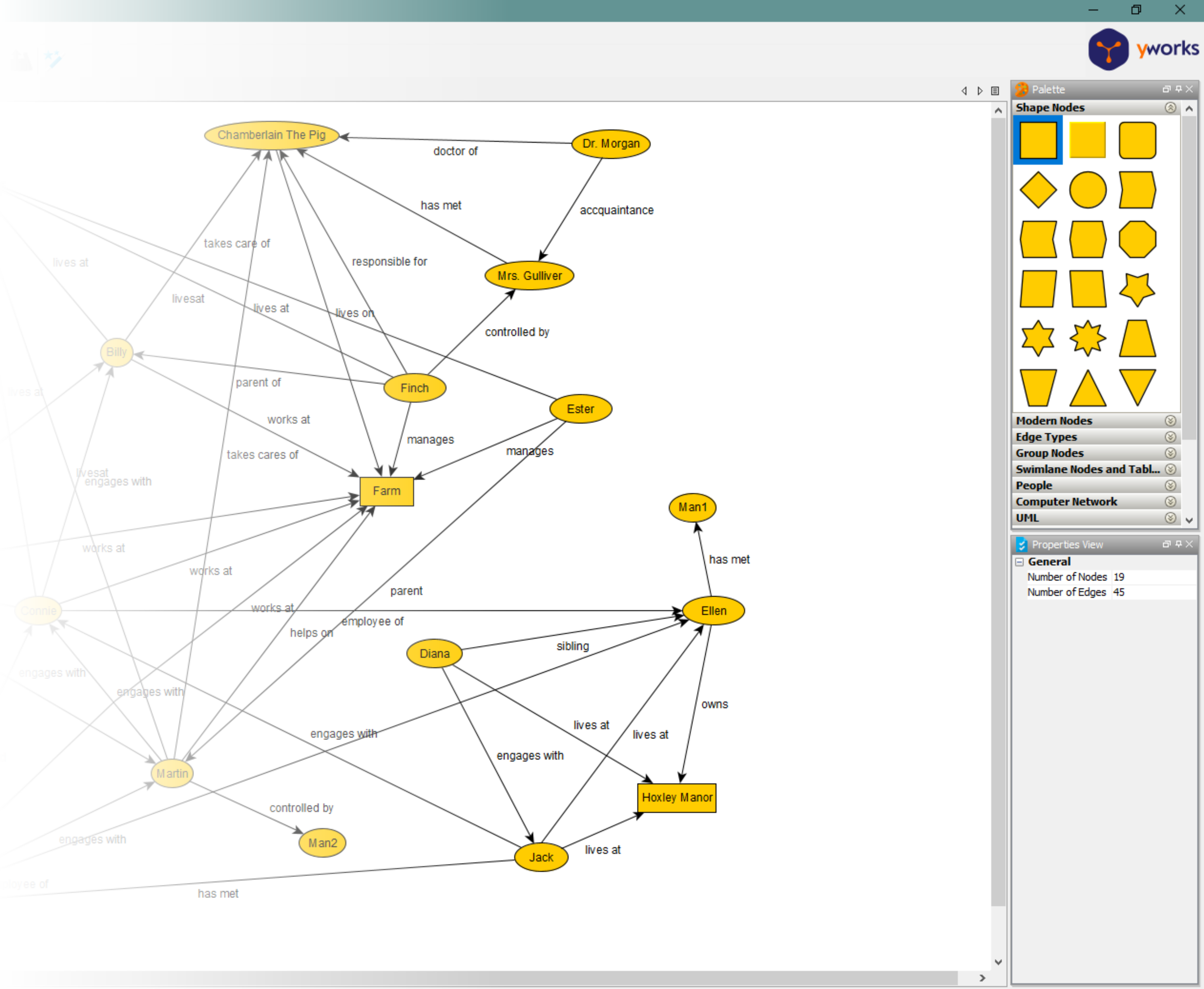
- Movies are first divided into scenes.
- A set of dedicated annotators were hired to work with us on the annotation framework[1].
- Annotators watch full movies, isolate and take images of main characters, places, & concepts. Draw Knowledge Graph (KG) of full movie using yEd* graphing tool.
- Annotators watch individual scenes, and draw KG over the scene level recording location, 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 entities (characters, locations)
- XGML graph file is processed later for query generation

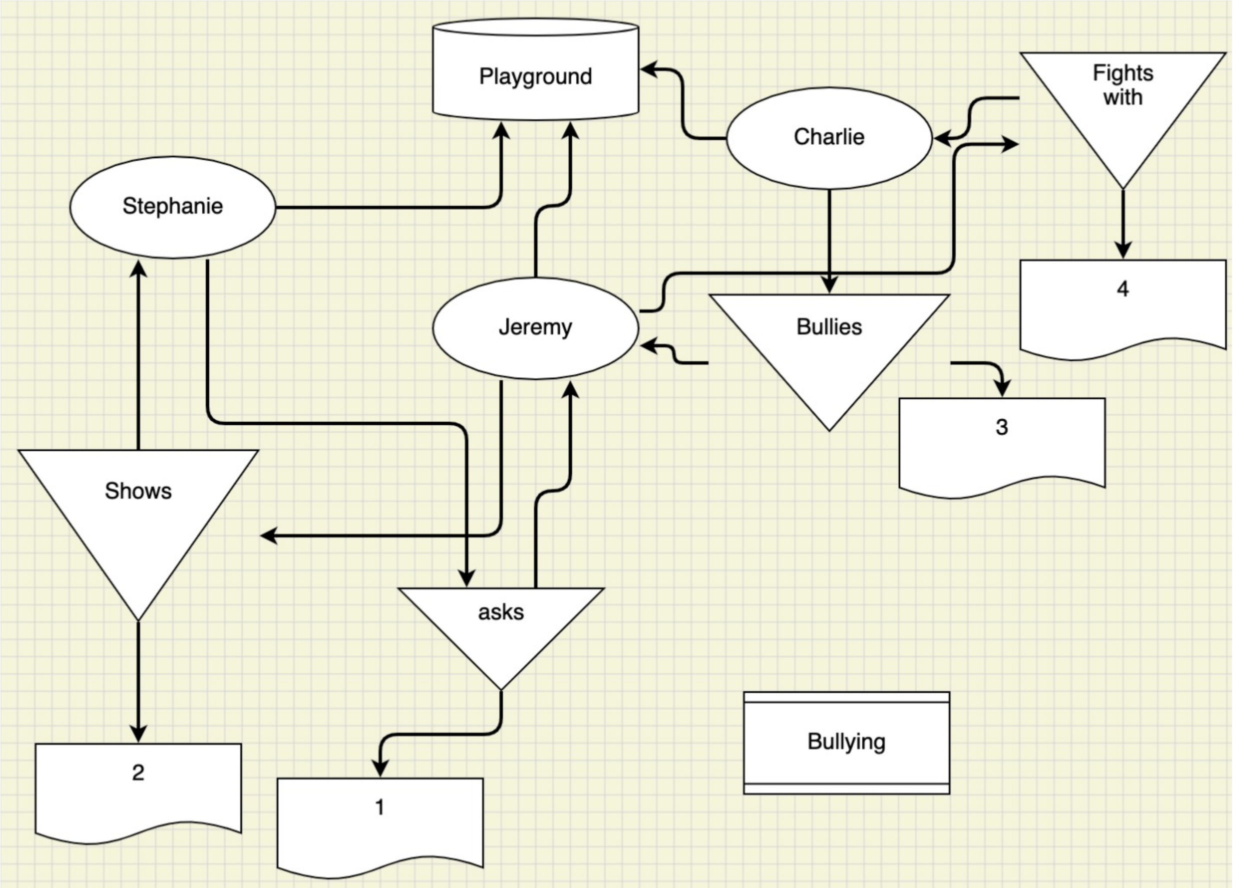


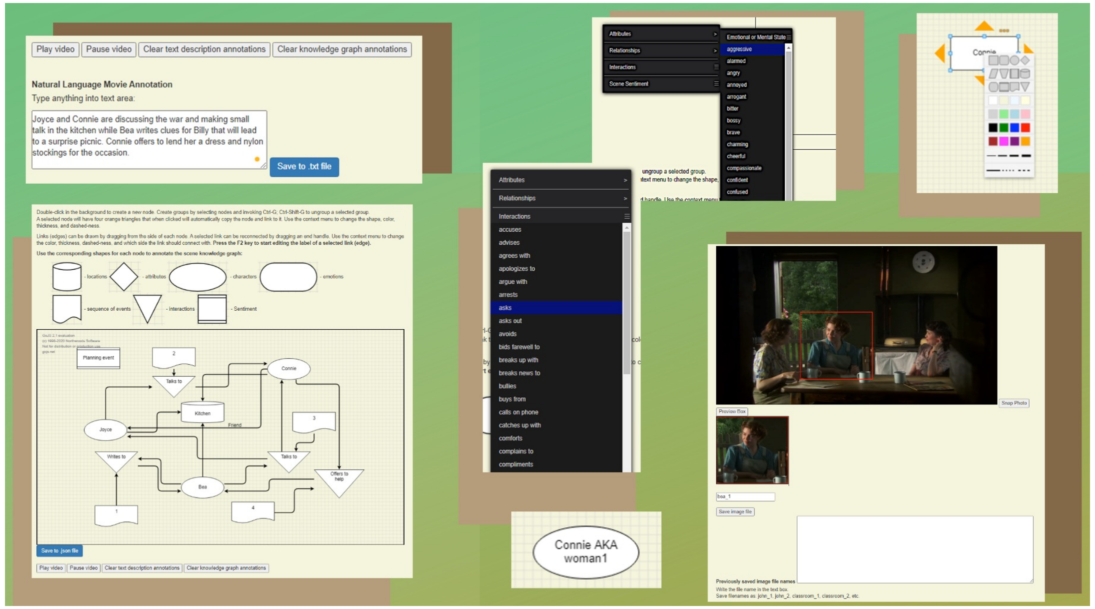
General

Number of Nodes	19
Number of Edges	45

Annotation: Scene-level

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





Natural Language Movie Annotation

Type anything into text area:

Joyce and Connie are discussing the war and making small talk in the kitchen while Bea writes clues for Billy that will lead to a surprise picnic. Connie offers to lend her a dress and nylon stockings for the occasion.

Double-click in the background to create a new node. Create groups by selecting nodes and invoking Ctrl-G; Ctrl-Shift-G to ungroup a selected group.
A selected node will have four orange triangles that when clicked will automatically copy the node and link to it. Use the context menu to change the shape, color, thickness, and dashed-ness.

Links (edges) can be drawn by dragging from the side of each node. A selected link can be reconnected by dragging an end handle. Use the context menu to change the color, thickness, dashed-ness, and which side the link should connect with. **Press the F2 key to start editing the label of a selected link (edge).**

Use the corresponding shapes for each node to annotate the scene knowledge graph:

- locations
- attributes
- characters
- emotions
- sequence of events
- interactions
- Sentiment

GoJS 2.1 evaluation
(c) 1998-2020 Northwoods Software
Not for distribution or production use
gojs.net

bea_1

Previously saved image file names

Write the file name in the text box.
Save filenames as: john_1, john_2, classroom_1, classroom_2, etc.

Queries: Movie-level

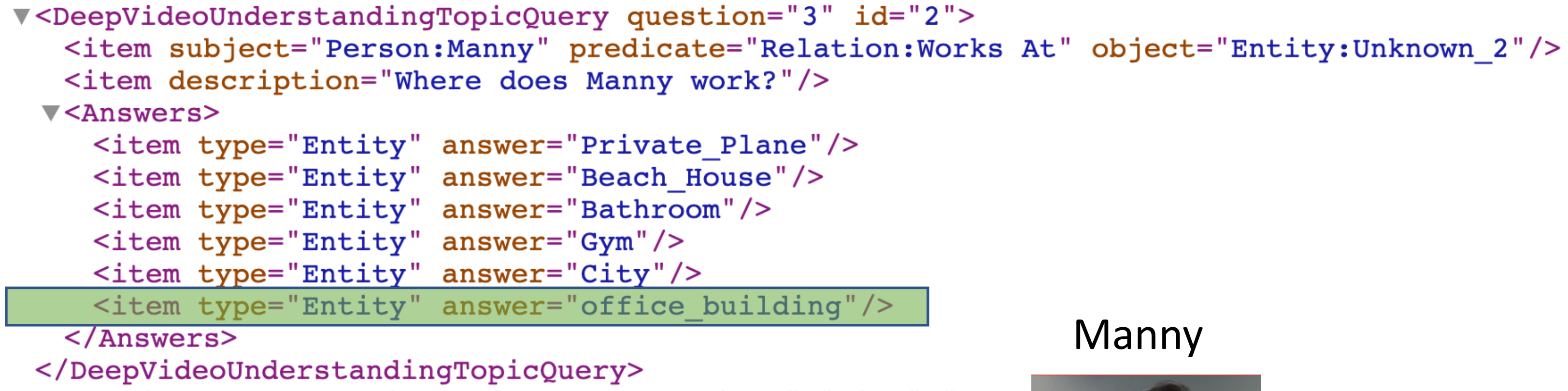
- Fill in the graph space: Given a list of entities, and/or relationships for certain nodes, where some nodes are replaced by variables X, Y, etc., solve for X, Y etc.

- Question Answering: This query type represents questions on the resulting KG in the form of multiple-choice questions. These queries also contain human-generated questions. These are open domain questions which are not limited to the ontology.

Queries: Scene-level

- Group 1:
  - Find the unique scene: Given a full, inclusive list of interactions, unique to a specific scene in the movie, teams should find which scene this is.
  - Find the next or previous interaction: 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*?
- Group 2:
  - Match the scene & text description: Given text descriptions and a list of scene numbers, match the correct scene numbers with text descriptions.
  - Scene sentiment classification: Given a scene number and a list of sentiment labels, which sentiment label belongs to that scene?

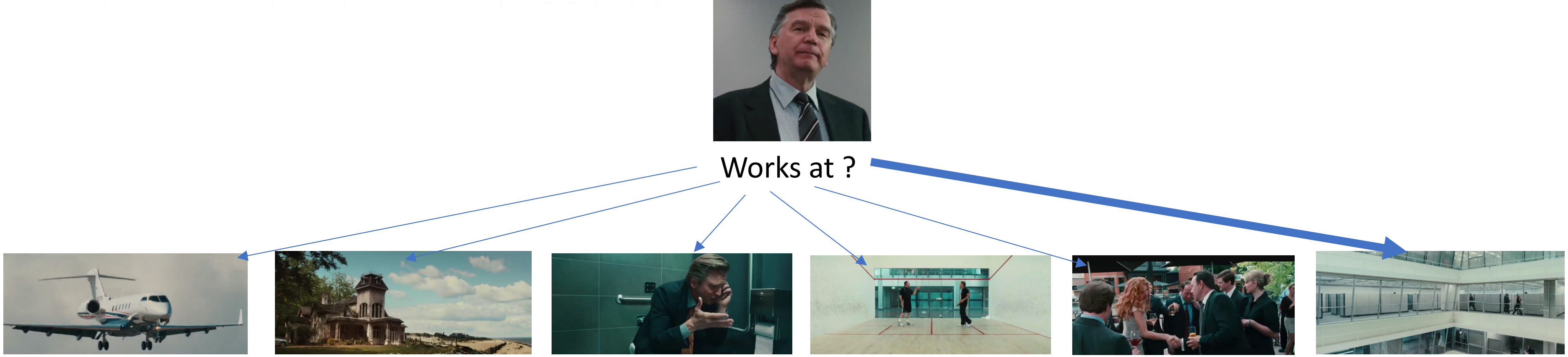
Query Samples: Movie-level



```
▼<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>
```

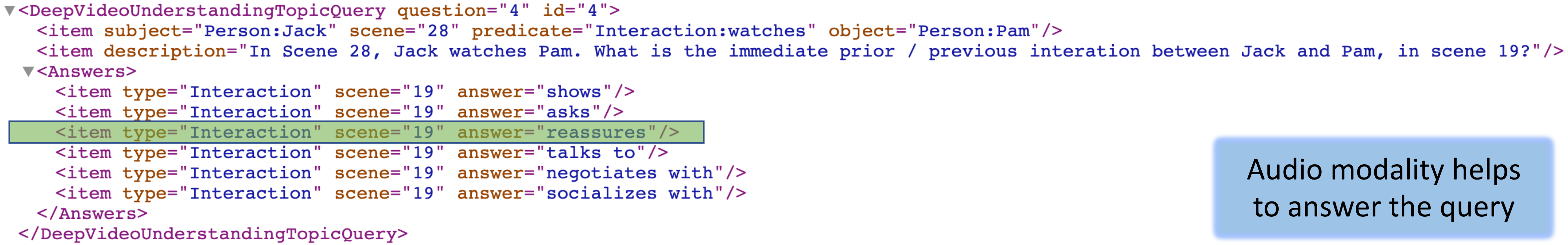
Visual modality helps to answer the query

Manny



**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 : Mean Reciprocal Rank (MMR).

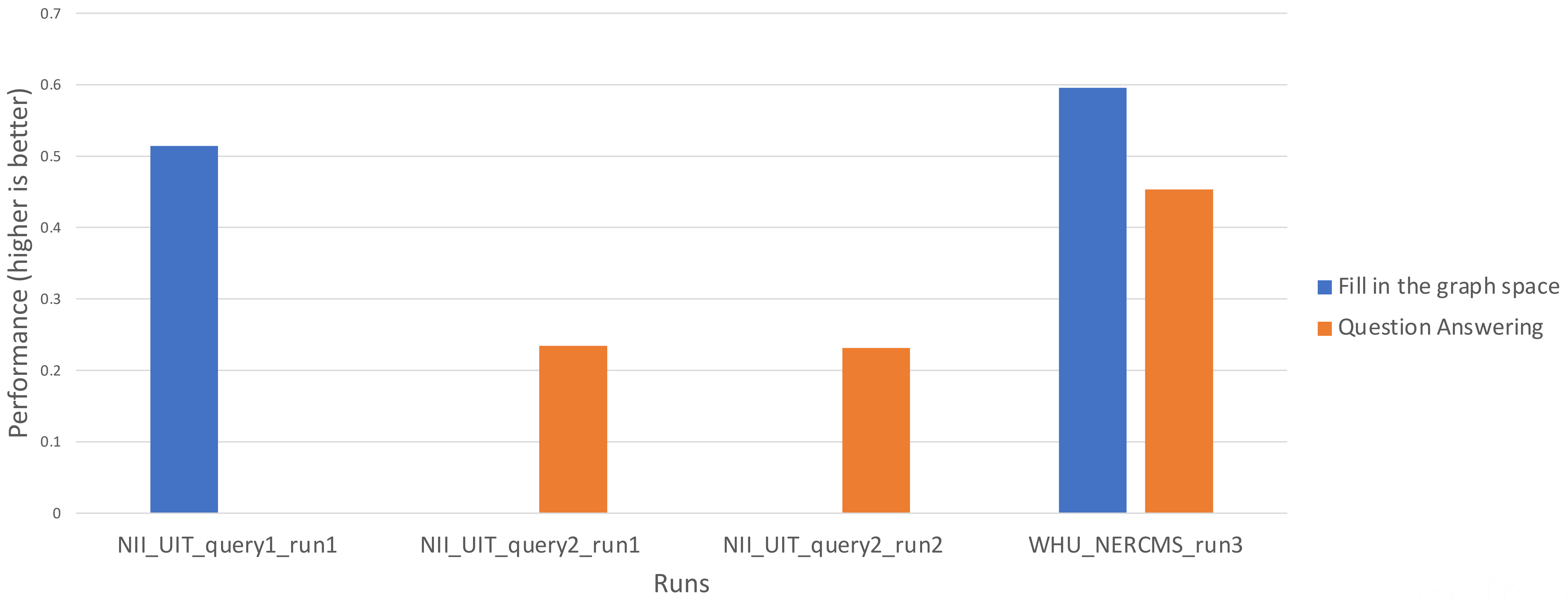
- Scene-Level
    - Next / Previous interaction : correct answers/total questions.
    - Find unique scene : Mean Reciprocal Rank (MMR).
    - Match descriptions to scenes: correct answers/total questions.
    - Scene sentiment classification : correct answers/total questions.

DVU 2023:

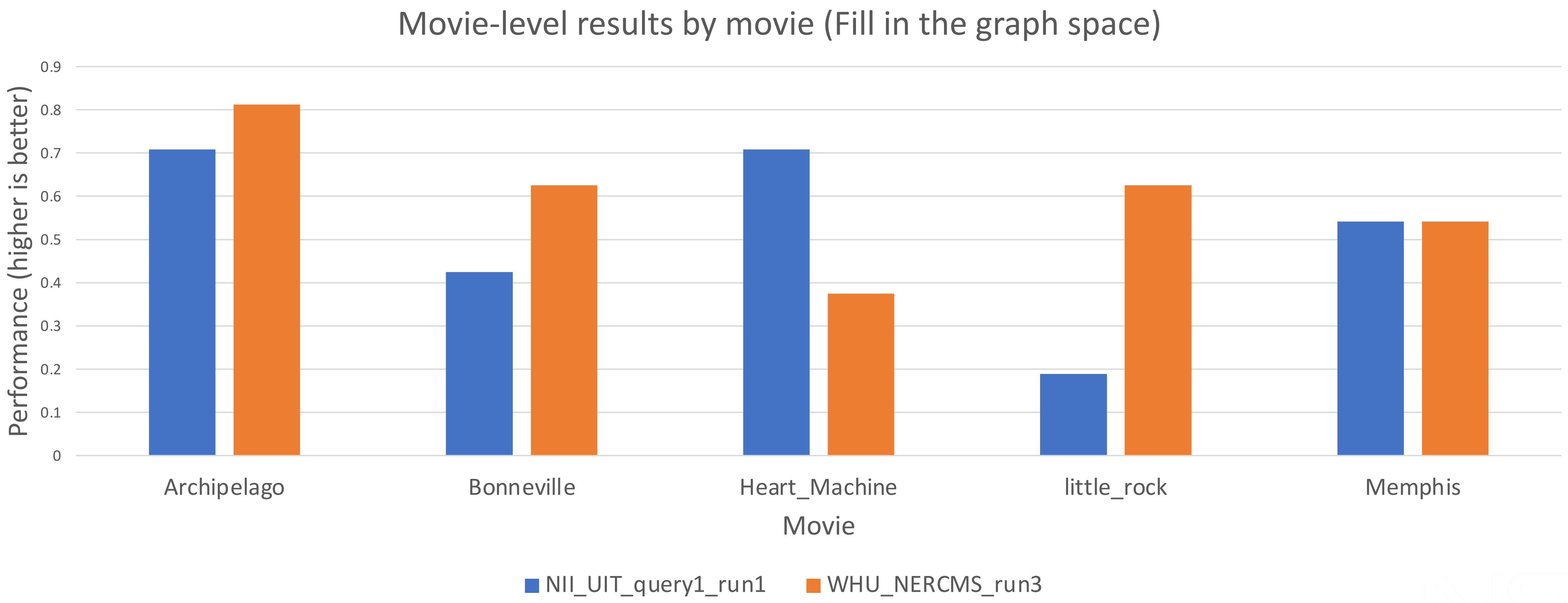
2 Finishers (out of 5 teams)

TEAM	ORGANIZATION	MOVIE-LEVEL RUNS	SCENE-LEVEL RUNS
NII_UIT	National Institute of Informatics, Japan; University of Information Technology, VNU-HCM, Vietnam	2	1
WHU_NERCMS	National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan City, Hubei Province, China	1	2

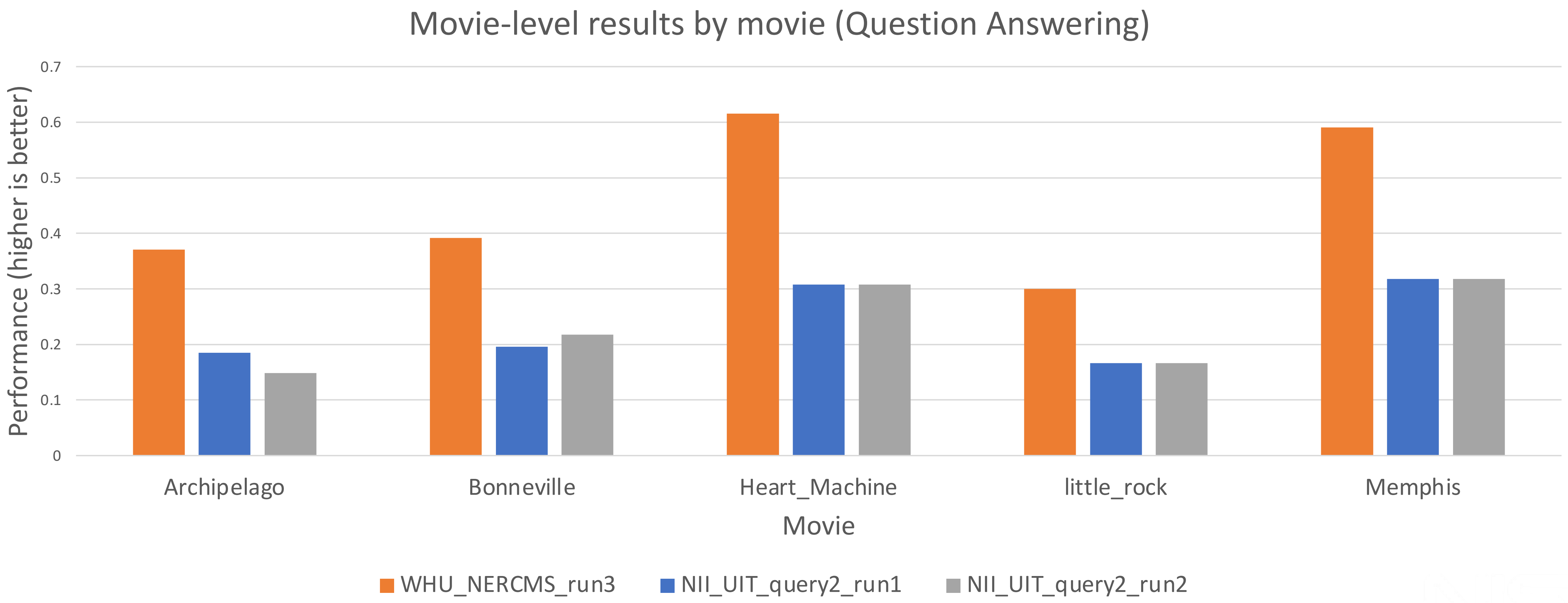
Movie-level Results (by run)



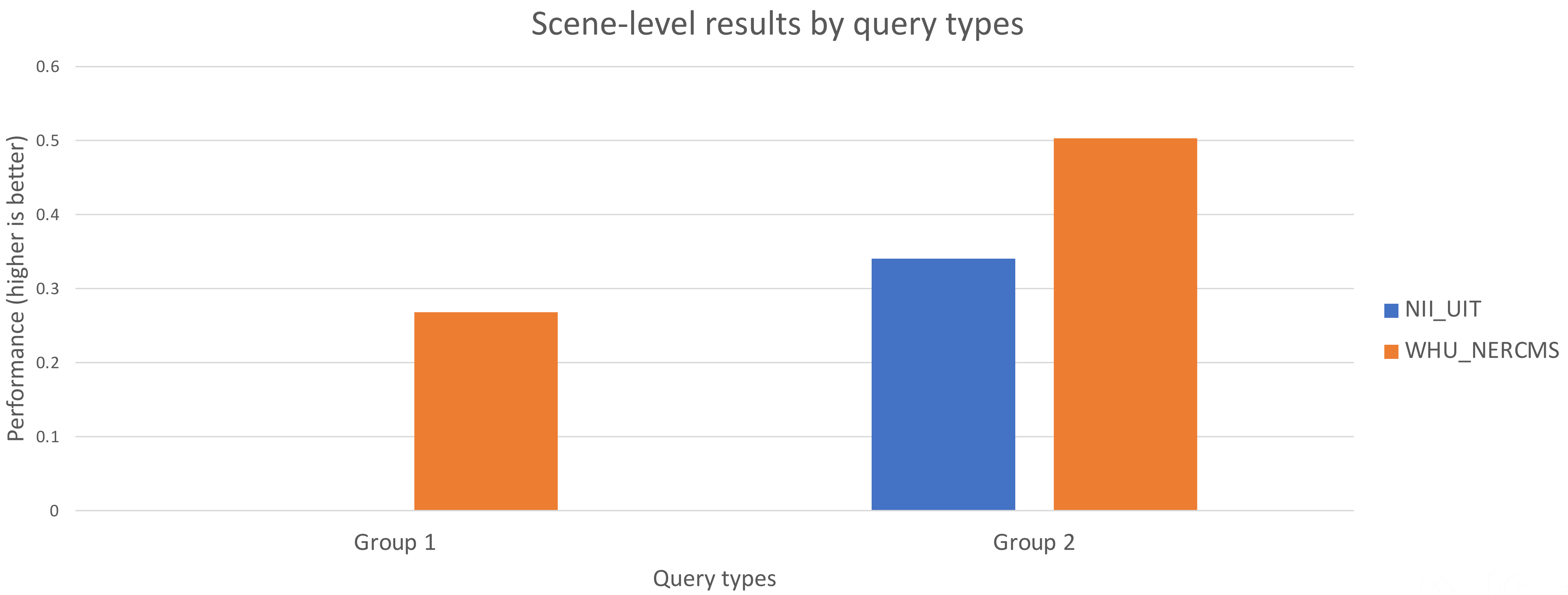
Results by query types : Movie-Level



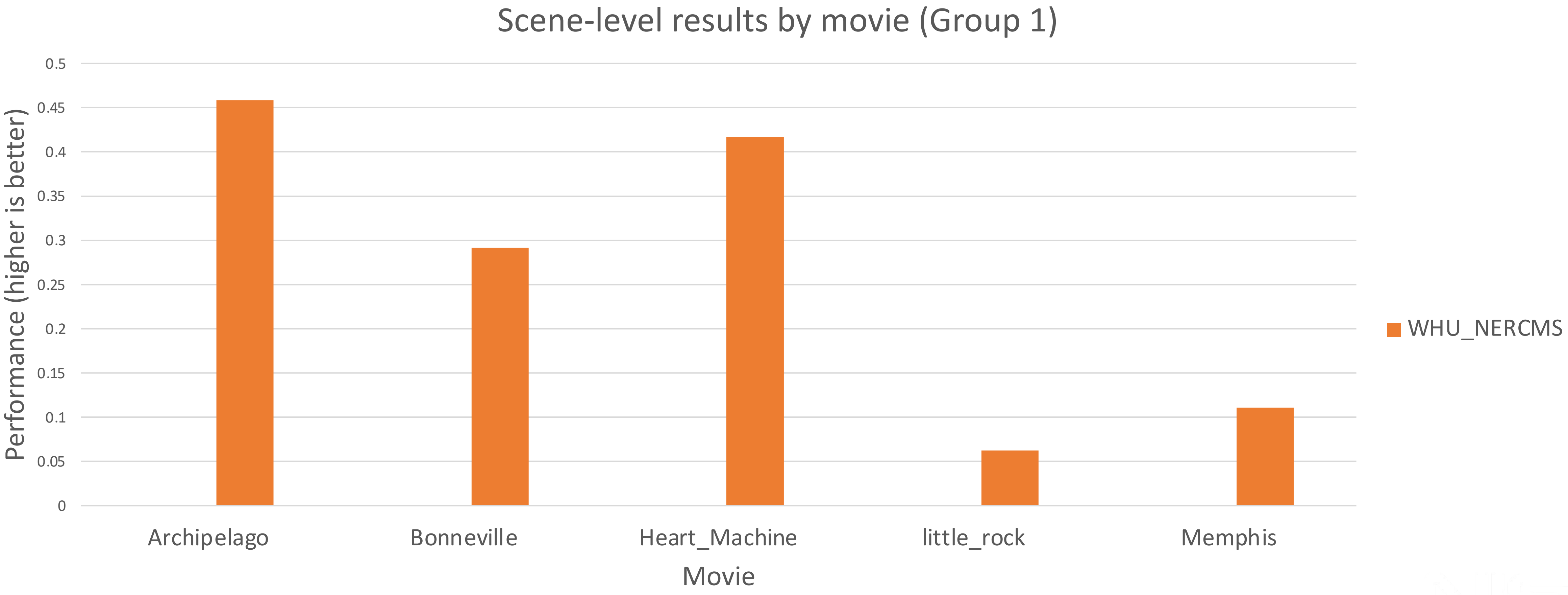
Results by query types : Movie-Level


Movie-level results by movie (Question Answering)

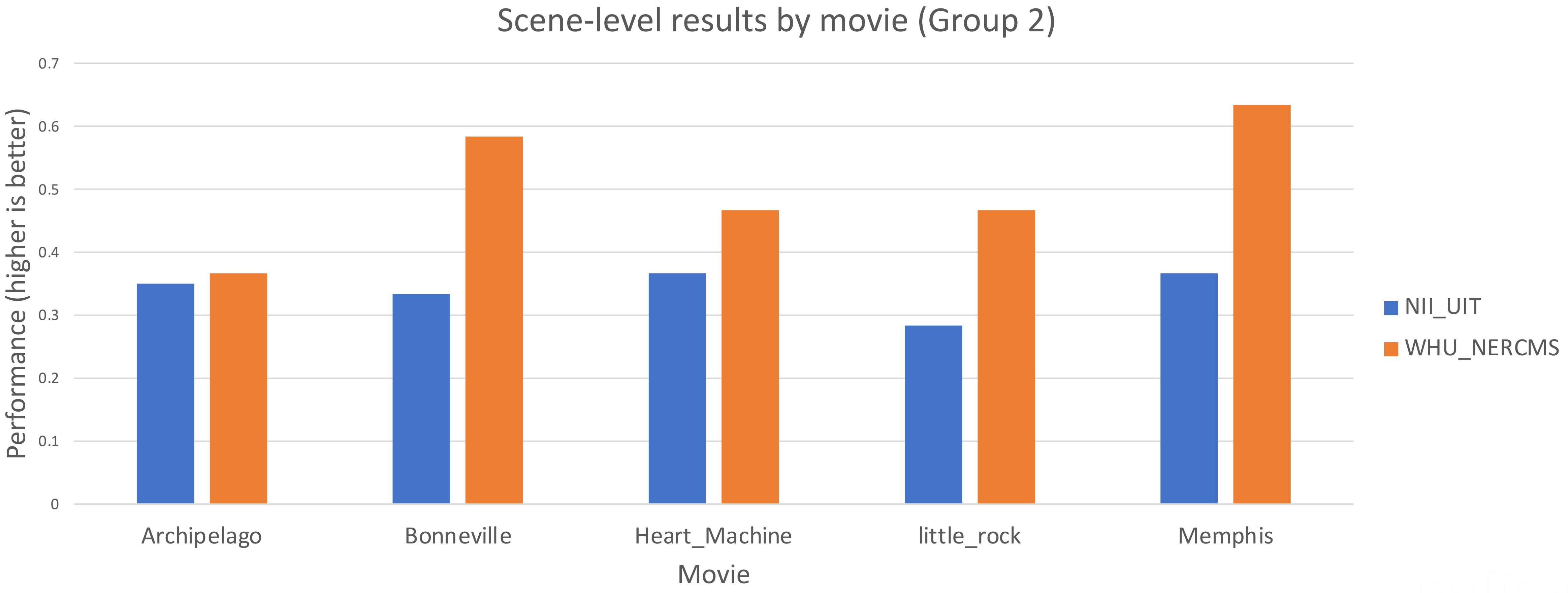
Scene-level Results



Results by query types : Scene-Level



Results by query types : Scene-Level



Conclusions

- Task participation is low (2 out of 5 teams finished).
- Movie-level fill in the graph space queries scored higher than question answering queries indicating QA queries are hard.
- Top system is consistently higher across most movies.
- Performance varies by movie.
- Scene-level group 2 queries (scene to text matching and sentiment classification) scored higher than group 1 queries (interactions focused).
- Overall movie-level results performed higher than scene-level results.
- LLMs are being applied to answer DVU queries.
- Given the low participation, the continuation of the task may not be feasible.
- We should target new extension tasks focused on multimodal understanding of long videos.