

TutoAI: a Cross-domain Framework for AI-assisted Mixed-media Tutorial Creation on Physical Tasks

Yuexi Chen¹, Vlad Morariu², Anh Truong², Zhicheng Liu¹



¹University of Maryland, College Park

² Adobe Research

Contact: ychen151@umd.edu



DEPARTMENT OF
COMPUTER SCIENCE



Adobe

How-to videos are popular

How-to videos are popular



How-to videos earn the **most attention** of any content category on YouTube, even more than music clips or gaming

Think with Google

Google/Ipsos, U.S., Video Mobile Diary (n of 18,219 total video occasions), 2017.

How-to videos are popular

Make pancakes



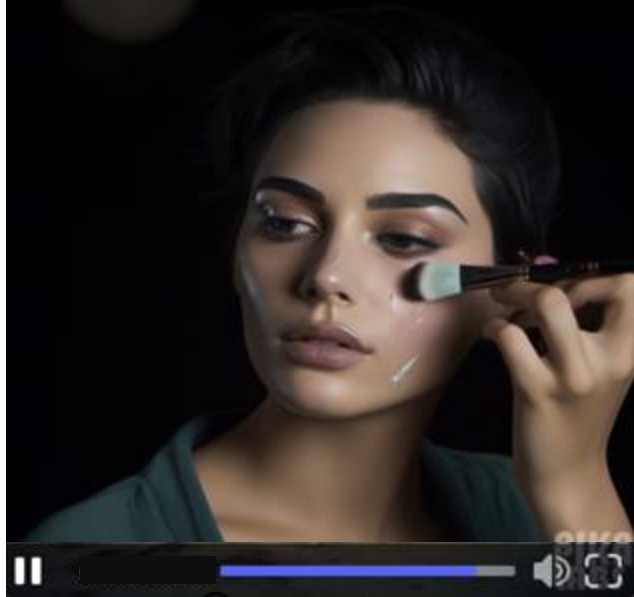
Apply makeup



Repair cars



Pain points of watching how-to videos



Lack of information overview



Timeline scrubbing is imprecise & tedious

Solution: mixed-media tutorials

Solution: mixed-media tutorials

[illegible]

The figure displays the user interface of a recipe recommendation system. It is divided into two main sections: a Video Section and a Control Section.

- Video Section:** Features a large video player showing a chef in a kitchen. Below the video is a progress bar and a search input field containing the text "beef (steak)". Underneath the search field are four small thumbnail images of different food dishes.
- Control Section:** Contains an "Actions" menu with a grid of filter buttons including "beef steak", "chicken breast", "beef", "steaks", "for two", "vegetarian", "dinner", "try new", "cheese", "lean protein", "quick", "easy", "vegetables", "salad", "buffet", "meatless week", "casserole", "party", "keto", "comfort", "break", "midweek", "spring forward", "vegan", "grain-free", "alcohol", "mild spiciness", "mild and low sodium", "low", "seasoning", "recipe".
- Recipe List:** Below the actions menu is a scrollable list of recommended recipes, each with a title and a duration:
 - introducing recipe (0:00-0:00)
 - introducing ingredients (0:00-0:00)
 - heat up pan (0:00-0:00)
 - seasoning steak (0:00-0:00)
 - roast steak (1:13-1:13)
 - try onions (1:00-0:00)
 - move onions to grill (0:00-0:00)
 - move steak to grill (0:00-0:00)
 - put cheese on the steak (0:00-0:00)
 - slice steak with cheese (0:00-0:00)

Liu et al. ConceptScape: Collaborative concept mapping for video learning. CHI'18; Yang et al. Improving Video Interfaces by Presenting Informational Units of Videos. CHI EA'22

TEAL GLITTER SMOKEY EYE (requested by Jaclyn Hill!) | PatrickStarr

PatrickStarr
@PSTARRSNAPS

ONLY VIDEO! This look was requested by Jaclyn Hill & was an honour to film glitter and color! especially since my camera isn't as good! I hope I like the Color! 24 Hours I expect some amazing rps with this look on my insta! :)

Twitter: @PSTARRSNAPS

TEAL

1. Just went to the hair salon to get my hair styled for the event on the 1st
2. Got phone so I can get my apply the oil over the hair so
3. The groom take the new foundation by Mrs. Jacobs. And
4. Next the groom take my undercoat? tomorrow and
5. The groom take the new foundation by Mrs. Jacobs. And
6. Next time I take the medium lipstick a red of about. And so

POSS

1. Just in going to the salon to get my hair styled for the event on the 1st
2. Got phone so I can get my apply the oil over the hair so

POSS

1. Just went to the hair salon to get my hair styled for the event on the 1st
2. Got phone so I can get my apply the oil over the hair so
3. The groom take the new foundation by Mrs. Jacobs. And
4. Next the groom take my undercoat? tomorrow and
5. The groom take the new foundation by Mrs. Jacobs. And
6. Next time I take the medium lipstick a red of about. And so

POSS

1. Just in going to the salon to get my hair styled for the event on the 1st
2. Got phone so I can get my apply the oil over the hair so

[illegible]

Solution: mixed-media tutorials

The image displays a video player interface showing a lecture on "superimposing alternate worlds". The video frame shows a man in a brown shirt standing in front of a chalkboard. The chalkboard has some writing on it, including "superimposing alternate worlds" and "world". Below the video frame, there is a progress bar and a timestamp of 8:47 / 13:28. To the right of the video frame, there is a control section with a list of actions: "introducing recipe", "introducing ingredients", "start up plan", "processing check", "start check", "my address", "move sensors to initial", "move check to initial", "put cheese on the bread", and "give cheese with cheese".

The main content of the video is a conceptual diagram of a system architecture. The diagram is a flowchart with nodes and arrows. The nodes are: "superimposing alternate worlds", "is read before for the purpose of", "superimposing alternate worlds", "comfort", "game engine", "audio renderer", "output side", "audio display", "visual display", "input", "head tracker", "collision detection", "self motion", "user motives", "motion data", "motion through an interface", "causes discomfort", "the movement is tracked in the alternate world", "data transfer", "data to sensors", "force feedback", "specific feedback", "game engine", "audio renderer", "output side", "audio display", "visual display", "input", "head tracker", "collision detection", "self motion", "user motives", "motion data", "motion through an interface", "causes discomfort", "the movement is tracked in the alternate world", "data transfer", "data to sensors", "force feedback", "specific feedback".

The diagram illustrates the flow of data and control between various components of a system. The components are represented by boxes, and the relationships are shown by arrows. The diagram is organized into several layers, with the top layer containing the main goal "superimposing alternate worlds". This goal is linked to "is read before for the purpose of", which then points to "superimposing alternate worlds". This component is further linked to "comfort", which is linked to "game engine". The "game engine" is linked to "audio renderer", which is linked to "output side". The "output side" is linked to "audio display" and "visual display". The "audio display" is linked to "audio renderer". The "visual display" is linked to "output side". The "output side" is linked to "input", which is linked to "head tracker". The "head tracker" is linked to "collision detection". The "collision detection" is linked to "self motion". The "self motion" is linked to "user motives". The "user motives" is linked to "motion data", which is linked to "motion through an interface". The "motion through an interface" is linked to "causes discomfort", which is linked to "the movement is tracked in the alternate world". The "the movement is tracked in the alternate world" is linked to "data transfer", which is linked to "data to sensors". The "data to sensors" is linked to "force feedback", which is linked to "specific feedback".

Figure 1: Screenshot of the ConceptScape interface. The interface is divided into several sections. On the left, there is a video player showing a man in a kitchen. Above the video player is a 'Video Section' with a timeline. To the right of the video player is a 'Control Section' with a list of actions and their durations. Below the video player is a 'beef (steak)' section with four small images showing the cooking process. On the far left, there is a 'TEAL GLITTER SMOKEY EYE' section with a video player and a list of actions. On the far right, there is a 'beef (steak)' section with a list of actions and their durations.

Auto-generated mixed-media tutorials are of low-quality



YouTube

Search



DIY | How To Make A SeeSaw For Kids Simple And Easy | 5 Fun Outdoor Activity Ideas with SeeSaw



Phuong mehneh
692 subscribers

Subscribe

176



Share

Download

Clip

Save



10K views 2 years ago SAN FRANCISCO

A seesaw is a staple game in most parks and playground. It is so fun, especially when you have somebody to share with.

Hayden and Irvin have started to play together much more, and so I thought it would be nice if they have a seesaw in the backyard, so they could hang out whenever they want.

In this video, I will show you how I made a bumble bee seesaw. This is a fun and easy project. It only took me a day to finish. I hope you'll enjoy making this seesaw with me. ...more

Chapters

These chapters are auto-generated



Intro
0:00



DIY SeeSaw
0:46



DIY Tire
5:38



Wood Blocks
8:08



bumblebee seesaw
9:26

All

From Phuong mehneh

Playgrounds



Road Trip with Kids || Animal
Petting Zoo || Weekend Getaw...
Phuong mehneh
1.3K views · 2 years ago



DIY Kids Wooden Climber
(Pikler Triangle) || Montessori ...
Phuong mehneh
8.4K views · 3 years ago



HOW TO INSTALL SYNTHETIC
TURF ON DIRT : Part 2 - ...
Phuong mehneh
87 views · 2 years ago



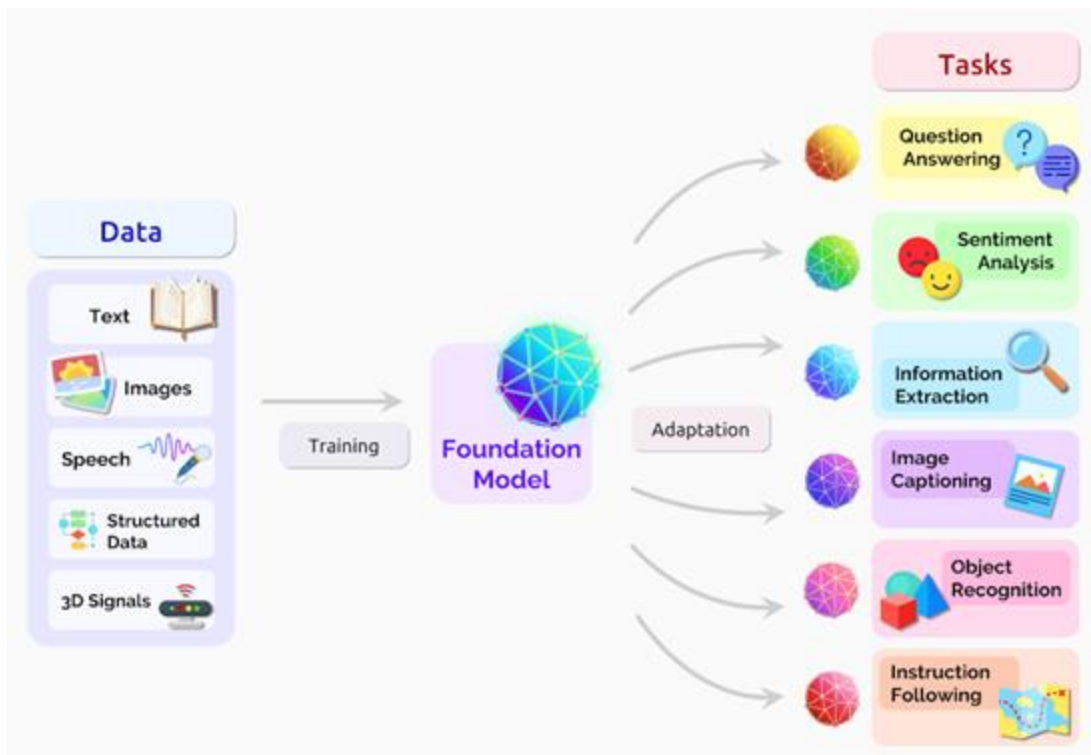
DIY BUSY BOARD | Cheap and
Easy Sensory Activity Idea To...
Phuong mehneh
4.5K views · 2 years ago

Creating high-quality tutorials require domain knowledge

For makeup tutorials,
cluster steps by facial parts



AI presents opportunities to create cross-domain tutorials



Challenges for AI-assisted cross-domain tutorial creation

Input: video, transcript

Output: mixed-media tutorials (videos, images, text, diagrams)

Challenges for AI-assisted cross-domain tutorial creation

Input: video, transcript

Output: mixed-media tutorials (videos, images, text, diagrams)

Challenges:

- **Vast model space:** multiple models are required and applicable
- **Inaccurate results:** models may make errors

Challenges for AI-assisted cross-domain tutorial creation

Input: video, transcript

Output: mixed-media tutorials (videos, images, text, diagrams)

Challenges:

- Vast model space: multiple models are required and applicable
- Inaccurate results: models may make errors

Research questions:

- How to **identify, evaluate, and select** models to create cross-domain mixed-media tutorials?
- How to design **user interfaces** to refine AI-generated mixed-media tutorials?

TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation

TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation

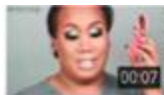
Components

Steps

text, video frames, timestamps

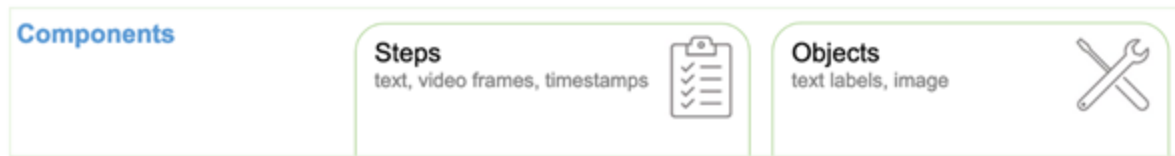


Examples:



24. I'm gonna take my favorite blush captivating by tarped.

TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation



Examples:



24. I'm gonna take my favorite blush captivating by tarped.

Things You'll Need

- ☐ Hose
- ☒ Roof cement
- ☐ Chisel
- ☐ Hammer

TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation

Components

Steps

text, video frames, timestamps



Objects

text labels, image



Dependencies

temporal order, spatial order



Examples:






24. I'm gonna take my favorite blush captivating by tarped.

Things You'll Need

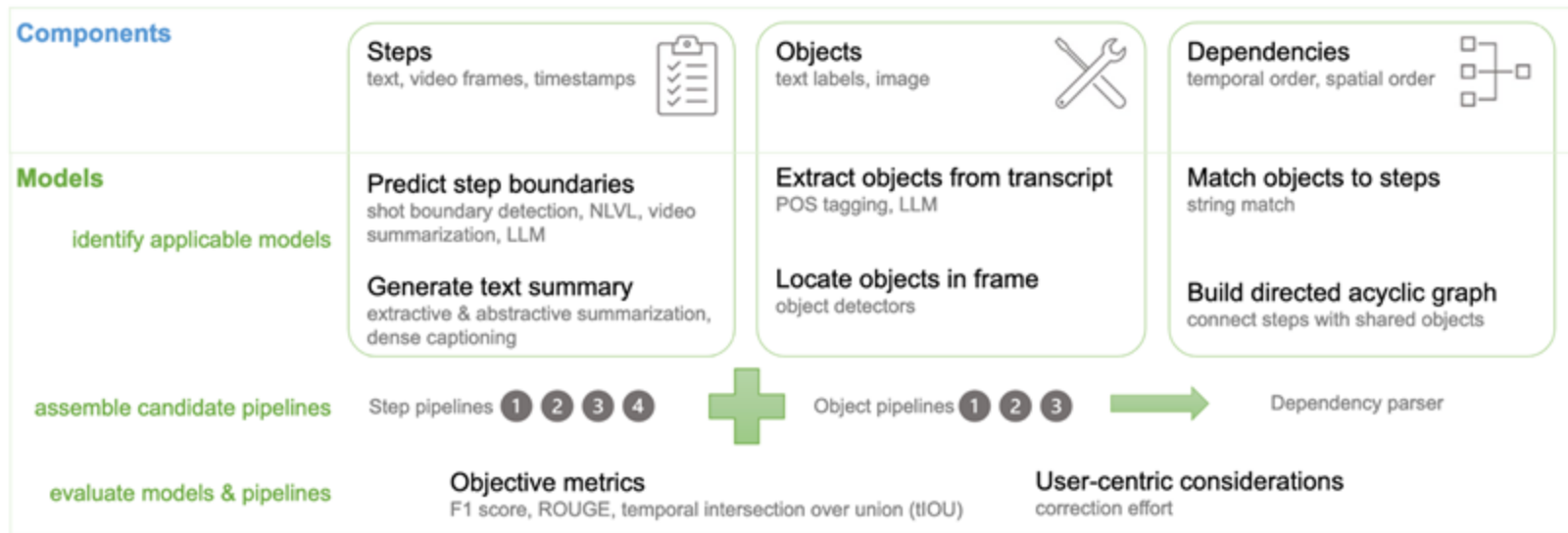
- ☐ Hose
- ☒ Roof cement
- ☐ Chisel
- ☐ Hammer



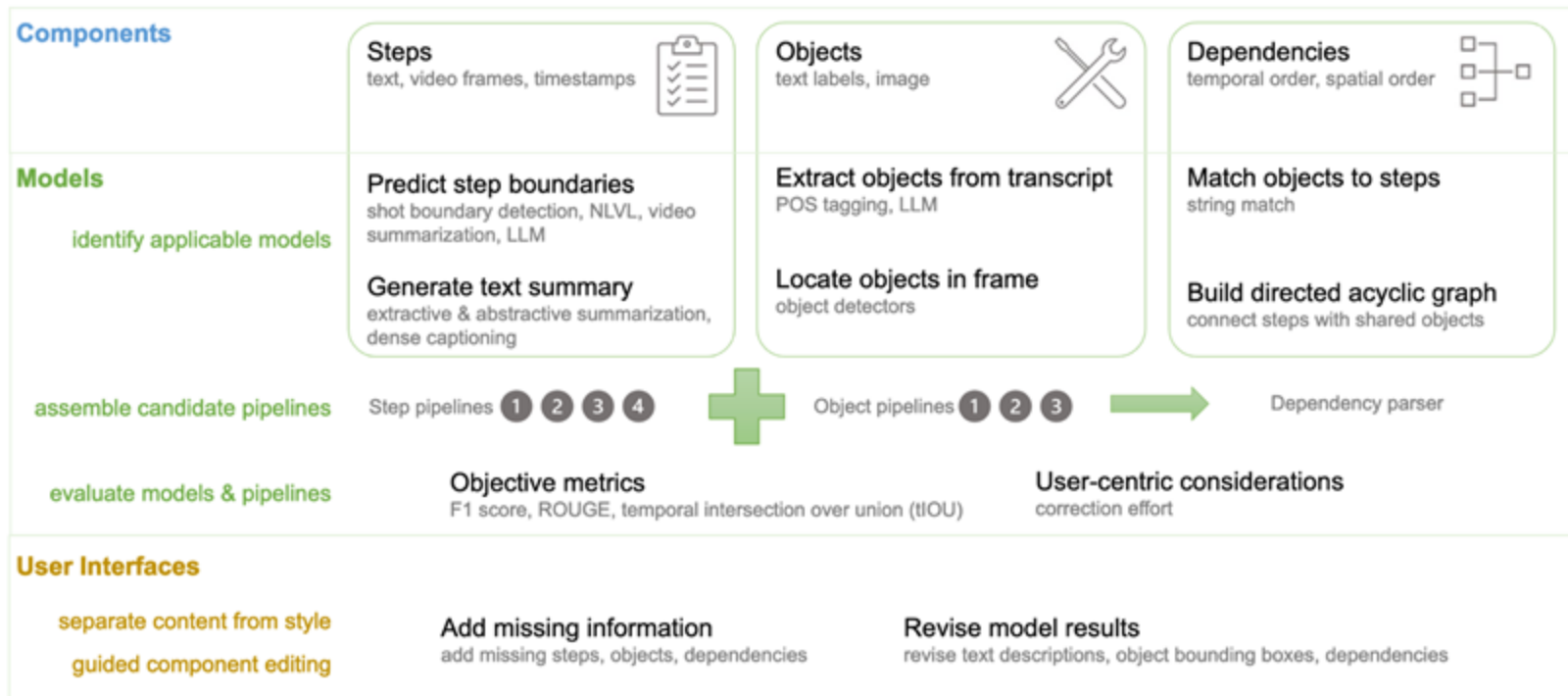
TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation

Components	Steps text, video frames, timestamps 	Objects text labels, image 	Dependencies temporal order, spatial order 
Models identify applicable models	Predict step boundaries shot boundary detection, NLVL, video summarization, LLM Generate text summary extractive & abstractive summarization, dense captioning	Extract objects from transcript POS tagging, LLM Locate objects in frame object detectors	Match objects to steps string match Build directed acyclic graph connect steps with shared objects

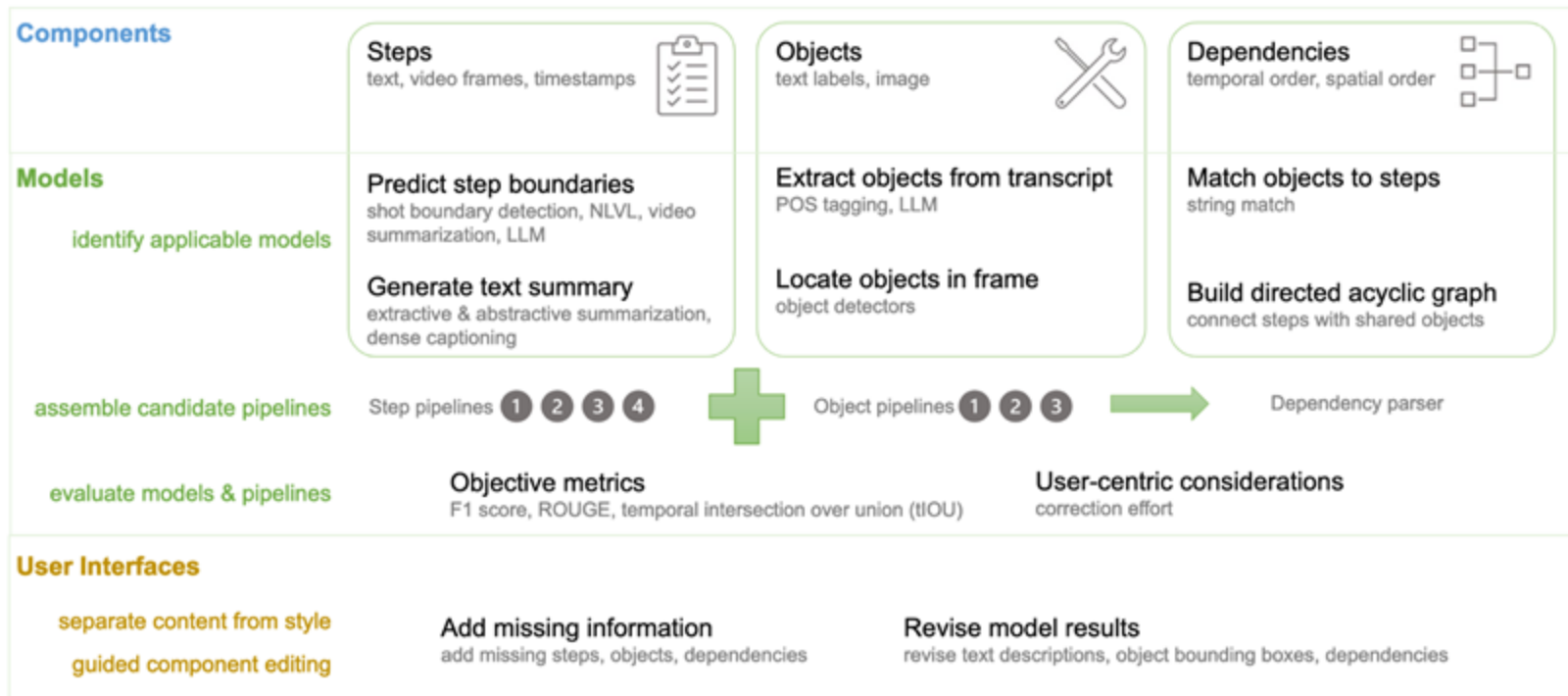
TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation



TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation



TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation



An object component example: “chicken breast”

“...here are the ingredients...this is the boneless skinless chicken breast thinly slice like this...”



<https://youtu.be/ntiGX3X-spA>

An object component example: “chicken breast”

*“...here are the ingredients...this is the boneless skinless **chicken breast** thinly slice like this...”*



<https://youtu.be/ntiGX3X-spA>

How to extract “chicken breast” from text and images?

Candidate ML pipelines for object extraction

#1



Pipeline #1: general object detectors



General object detectors (e.g., ResNet) are limited by the training dataset

Computation time on cpu: 1.228 s

dining table

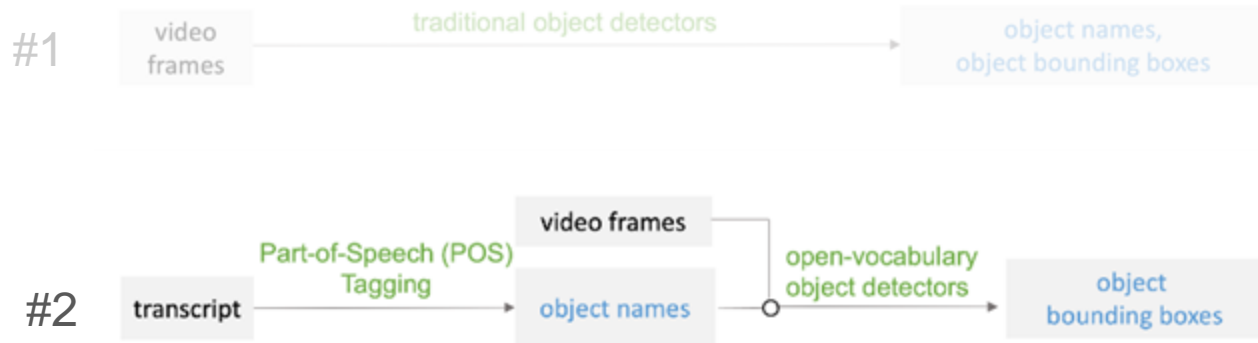
0.978

bowl

0.998

<https://huggingface.co/facebook/detr-resnet-50>

Candidate ML pipelines for object extraction

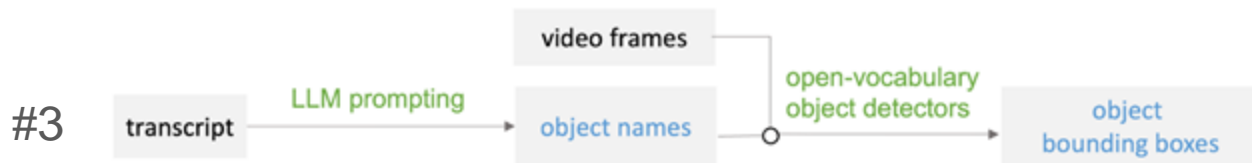
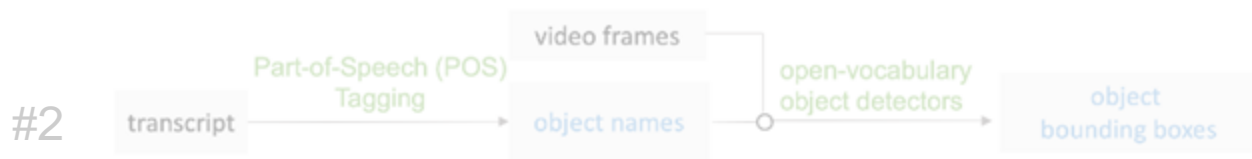


Pipeline #2: part-of-speech tagging

input: *"...here are the ingredients...this is the boneless skinless chicken breast thinly slice like this..."*

Output (only keep nouns): *"ingredients", "chicken", "breast"*

Candidate ML pipelines for object extraction



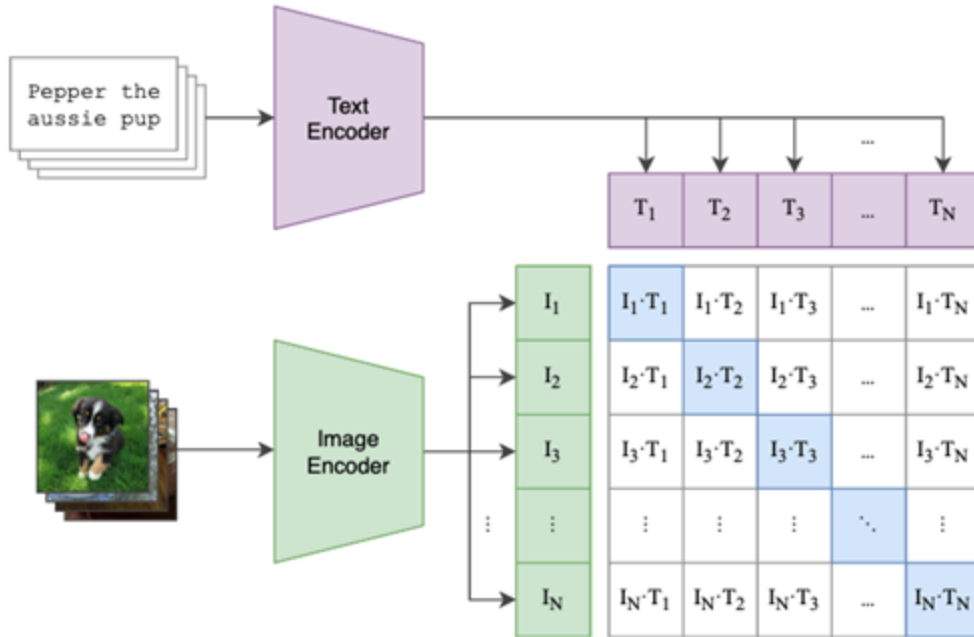
Pipeline #3: large language model (LLM) prompting

Input: the entire transcript

Prompt: “Identify the ingredients in this tutorial”

Output: “Boneless skinless chicken breast”, ...

Pipeline #3: detect objects given text labels



Radford et al. "Learning transferable visual models from natural language supervision." ICML 2021

Pipeline #3: open-vocabulary object detection

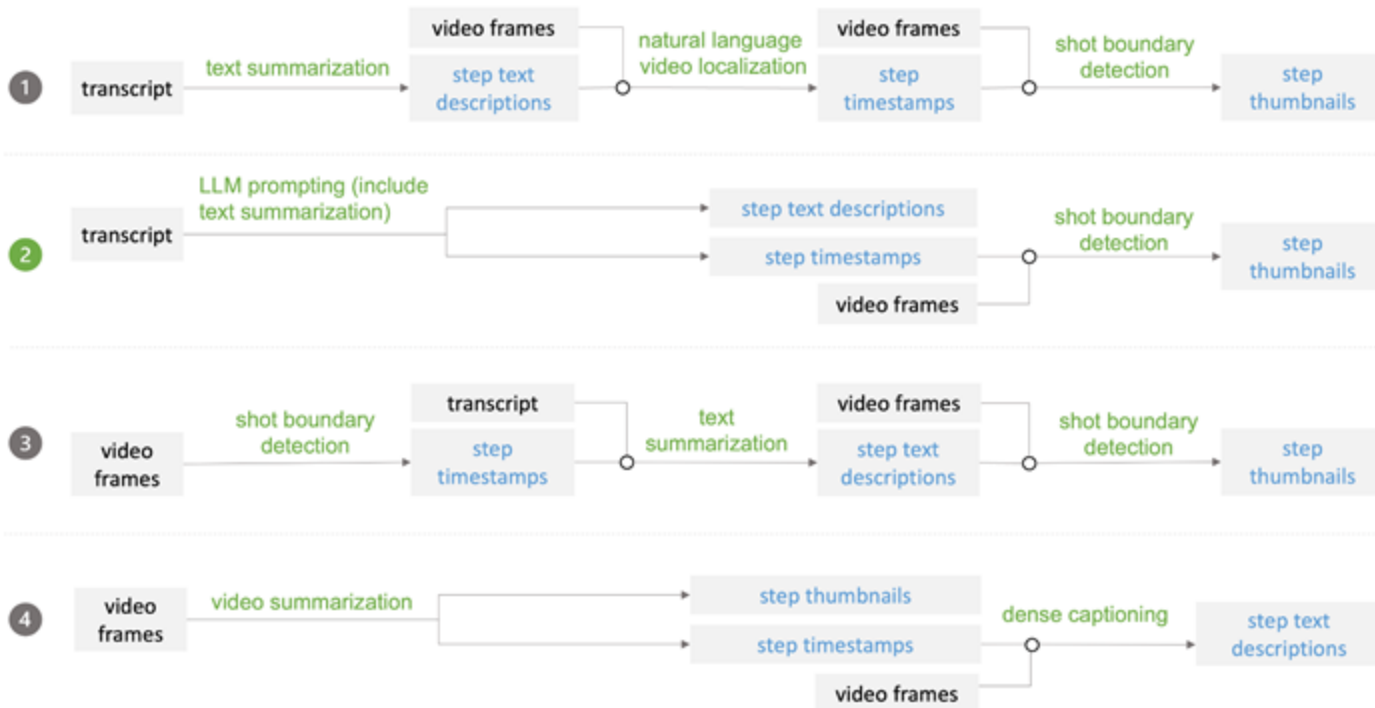


Labels extracted by LLM:

- Boneless skinless chicken breast
- Chopped green pepper
- Roasted peanuts
- Red chili
- Green onions

Candidate ML pipelines for step extraction

Step extraction



Evaluation: a large-scale cooking video dataset

Dataset:

- ❑ 347 annotated cooking videos
- ❑ 20 diverse instructional videos (crafting, makeup, repair)



Evaluation: a large-scale cooking video dataset

Dataset:

- ❑ 347 annotated cooking videos
- ❑ 20 diverse instructional videos (crafting, makeup, repair)



Metrics:

- ❑ Objective:
 - ❑ F1 scores
 - ❑ ROUGE scores
 - ❑ Temporal Intersection over Union (tIOU)
- ❑ Subjective:
 - ❑ correction efforts required from humans

Edit

View

Make a seesaw for kids



IDENTIFY STEPS



CHOOSE THUMBNAILS



SELECT OBJECTS

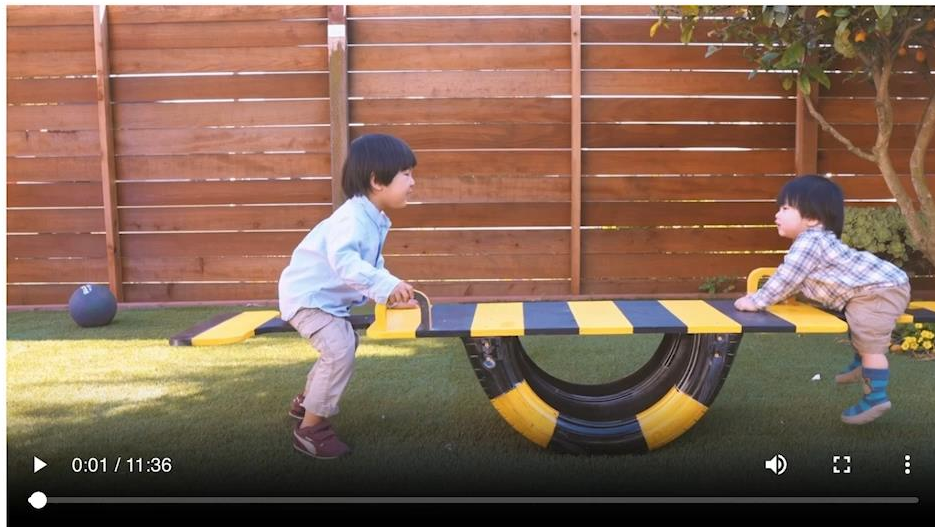


CROP OBJECTS



BUILD DEPENDENCIES

Instructions



Step #: 1 < Snap Start > Snap End



- 00:02 a seesaw is a staple game in most parks and playground
- 00:06 it is so fun especially
- 00:08 when you have somebody to share with irvine and hayden have started to play together much more
- 00:13 so i thought it would be so fun if
- 00:16 they have a seesaw in the backvard so tehy could hand out whenever they

1 00:01 - 00:19



Introduction to the project and purpose of creating a bumblebee seesaw



2 00:20 - 01:22



Gathering materials including a used tire and plywood board



3 01:23 - 02:54



Measuring and marking the board for the curved seating areas



4 02:55 - 03:38



Cutting and sanding the board



5 03:39 - 06:31



Painting the board with black and yellow stripes using painter's tape



6 06:32 - 08:02



Cutting the tire in half and priming and painting it with black and yellow stripes



7 08:03 - 09:06



Creating wood blocks to attach the tire to the board



8 09:07 - 10:46



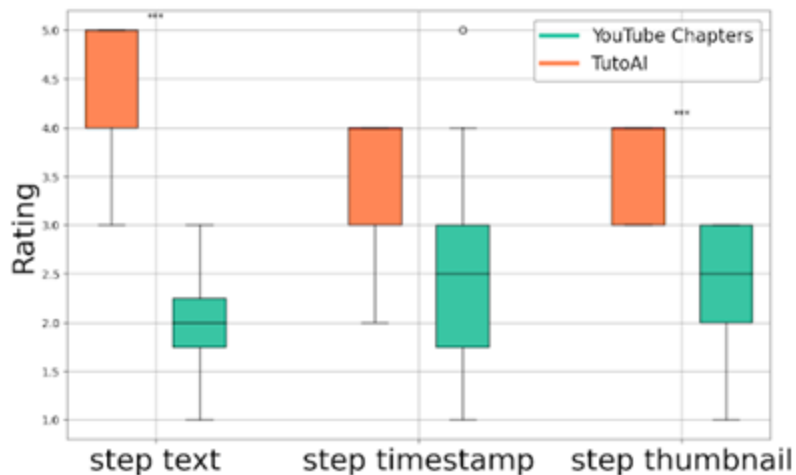
Attaching the tire and handles to the board



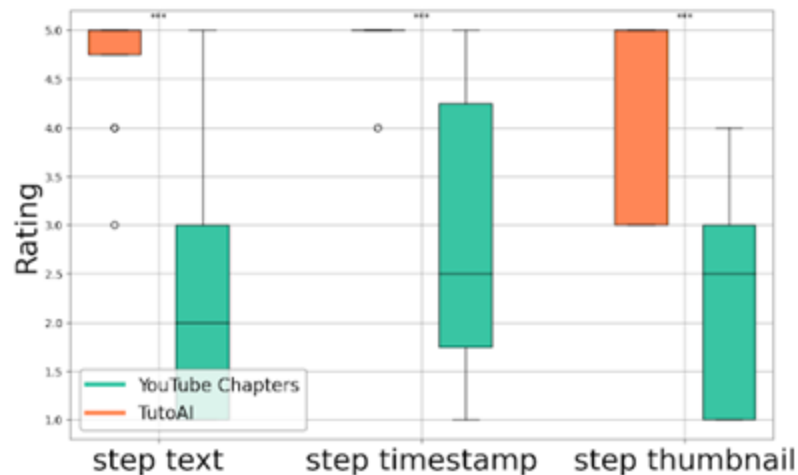
TutoAI vs. YouTube Chapters

TutoAI components received higher ratings than **YouTube** Chapters (**N=12**)

Component quality **before editing**



Component usefulness **after editing**



YouTuber's testimonial

*"I think this is a great tool. I like this a lot...this is giving me the ability to do a lot more, especially creating the **flow charts** ...viewers would get a lot out of this as opposed to just a regular chapter..."*



SevenFortyOne Radios and Repairs

@SevenFortyOne · 43.1K subscribers · 590 videos

Ham Radio demonstration and repair with some Chevy Suburban and home repairs mixed in... >

twitter.com/SevenFortyOne8 and 1 more link

Subscribe

TutoAI guidelines for human-centered system builders

TutoAI guidelines for human-centered system builders

Adopt a multi-modal perspective

👁 INITIALY —————

TutoAI guidelines for human-centered system builders

Adopt a multi-modal perspective

👁 INITIALY —————

Simplify complex creation by guiding and constraining user actions

Separate content from styles

🔗 DURING INTERACTION —————

TutoAI guidelines for human-centered system builders

Adopt a multi-modal perspective

👁️ INITIALLY —————

Focus on user-centric model selection

Support graceful degradation

⚠️ WHEN WRONG —————

Simplify complex creation by guiding and constraining user actions

Separate content from styles

🔗 DURING INTERACTION —————

TutoAI guidelines for human-centered system builders

Adopt a multi-modal perspective

👁 INITIALY —————

Focus on user-centric model selection

Support graceful degradation

⚠ WHEN WRONG —————

Simplify complex creation by guiding and constraining user actions

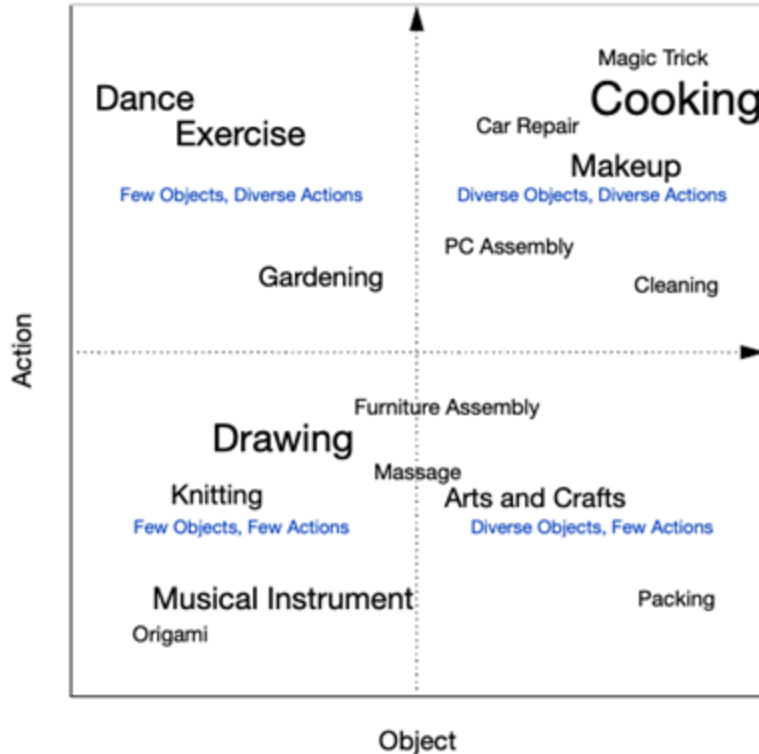
Separate content from styles

🔗 DURING INTERACTION —————

Leverage strong models for cross-modal enhancement

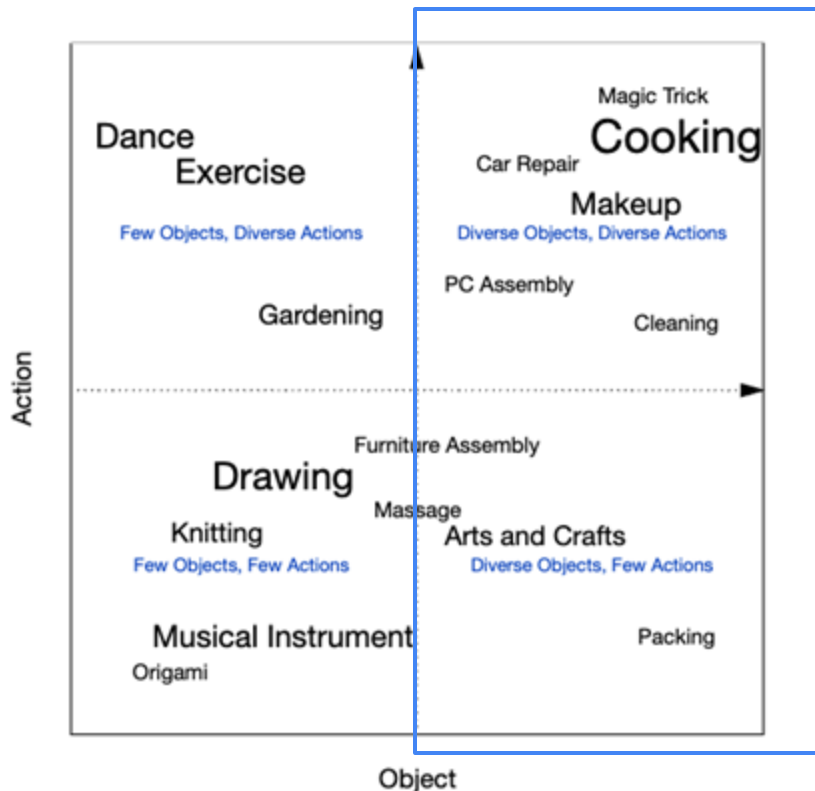
🕒 OVER TIME —————

Limitations and future work



Chang et al. "Rubyslippers: Supporting content-based voice navigation for how-to videos." CHI 2021.

Limitations and future work



Chang et al. "Rubyslippers: Supporting content-based voice navigation for how-to videos." CHI 2021.

Acknowledgment



Vlad I Morariu@Adobe



Anh Truong@Adobe



Leo Zhicheng Liu@Univ of Maryland



DEPARTMENT OF
COMPUTER SCIENCE

