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

Yuexi Chen[1], Vlad Morariu[2], Anh Truong[2], Zhicheng Liu[1]




[1]University of Maryland, College Park
[2] Adobe Research

Contact: ychen151@umd.edu

DEPARTMENT OF COMPUTER SCIENCE

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.

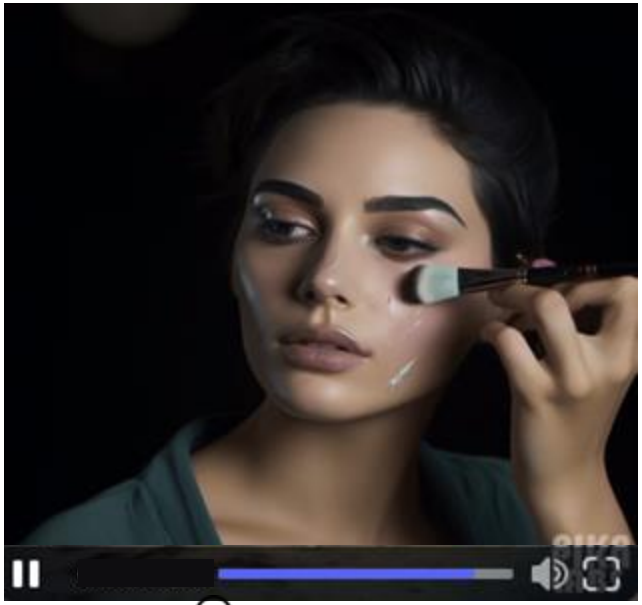
How-to videos are popular

Make pancakes

Apply makeup

Repair cars

Pain points of watching how-to videos

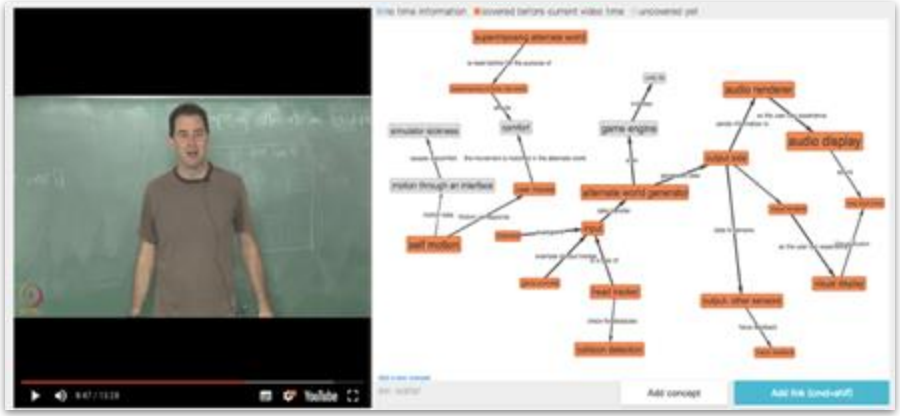


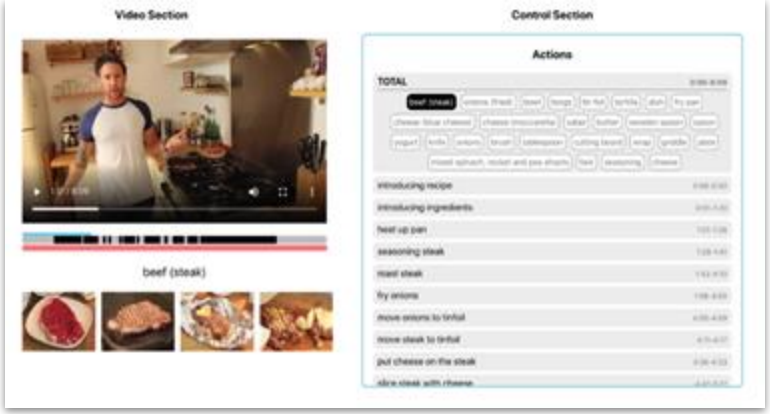
Lack of information overview

Timeline scrubbing is imprecise & tedious

Solution: mixed-media tutorials

Solution: mixed-media tutorials

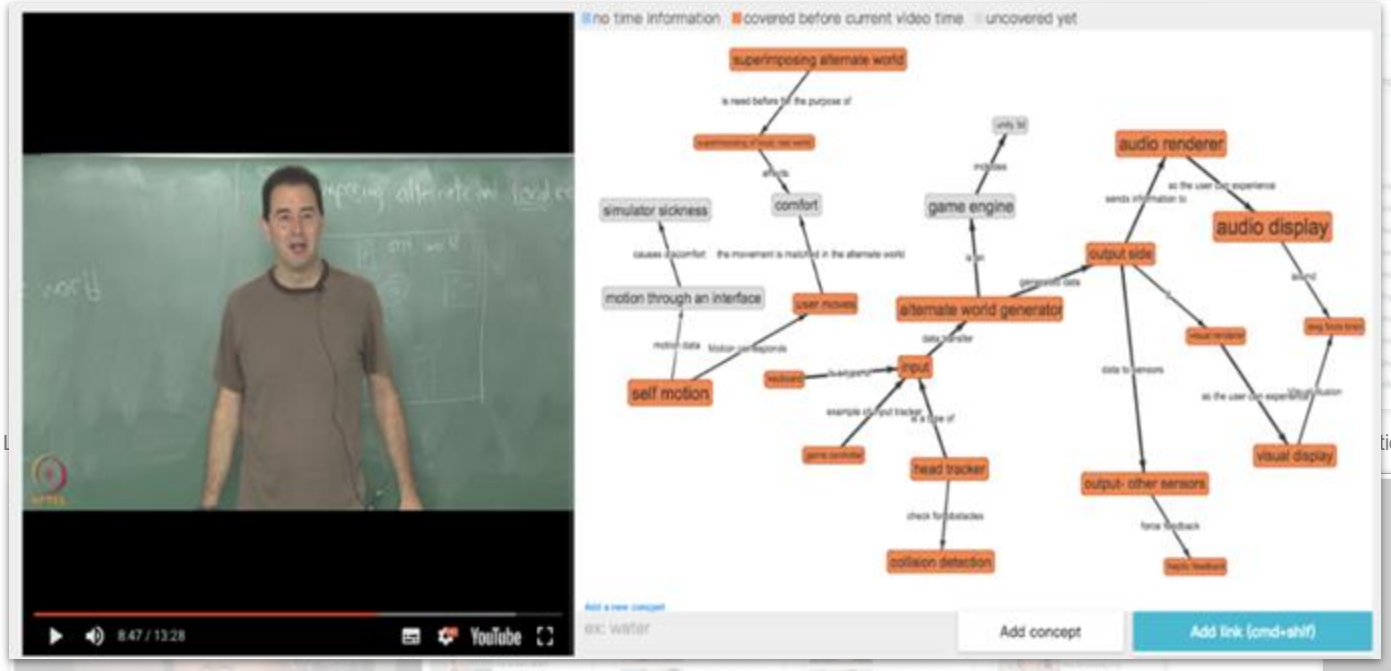


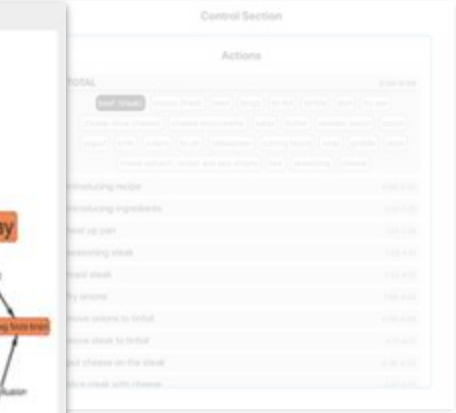


Liu et al. ConceptScape: Collaborative concept mapping for video learning. CHI'18; Yang et al. Improving Video Interfaces by Presenting Informational Units of Videos. CHI EA'22

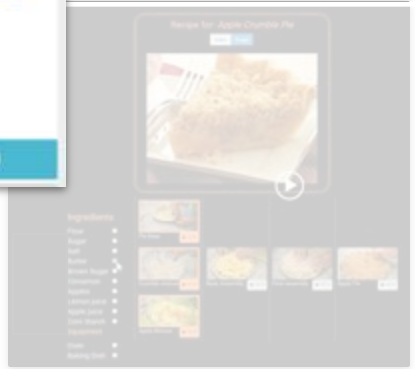
Truong et al. Automatic generation of two-level hierarchical tutorials from instructional makeup videos. CHI'21; Nawhal et al. VideoWhiz: Non-Linear Interactive Overviews for Recipe Videos. GI'19

Solution: mixed-media tutorials

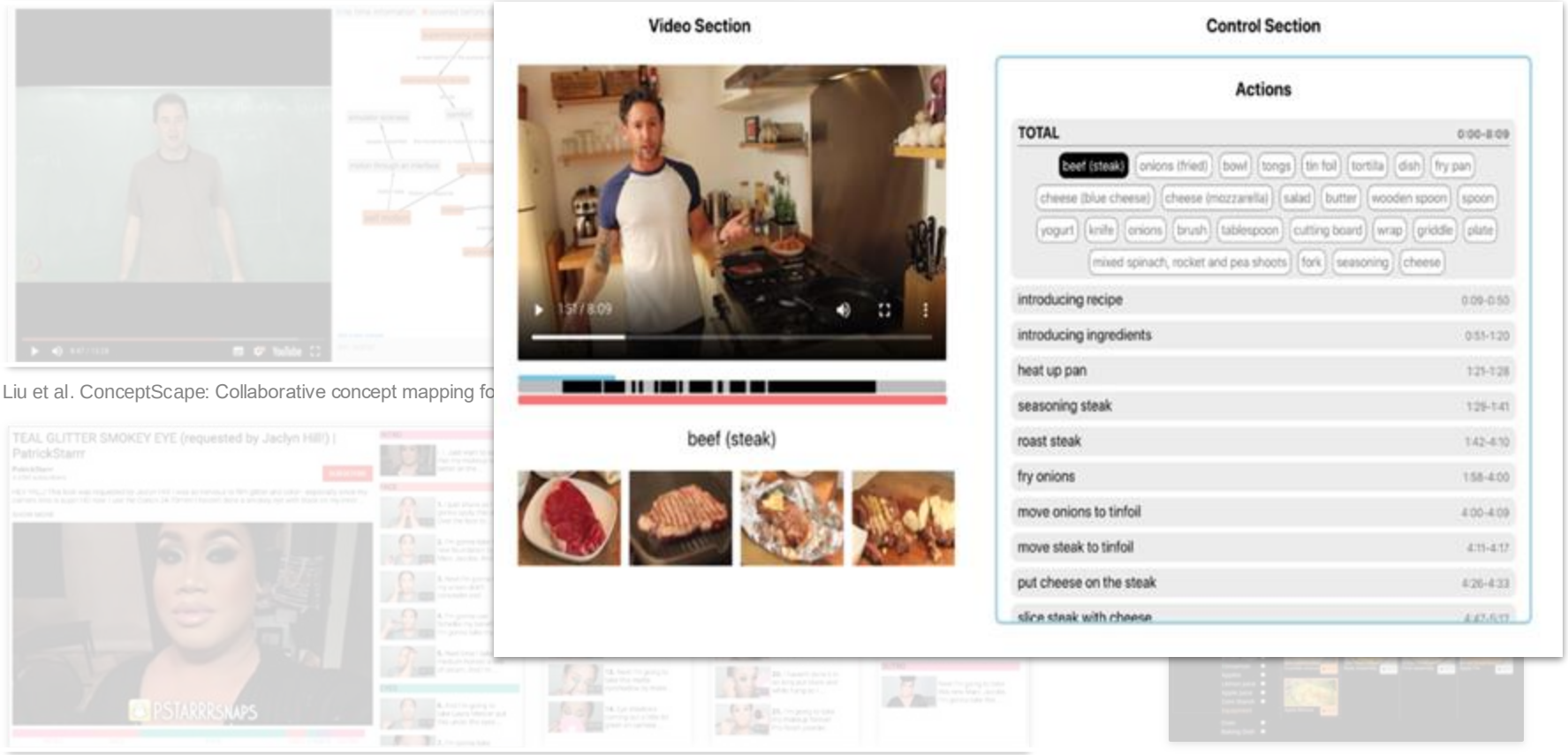




tional Units of Videos. CHI EA'22



Solution: mixed-media tutorials



Liu et al. ConceptScape: Collaborative concept mapping fo

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

https://youtu.be/drDSY3ZZqnQ

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

Phuong mehmeh
692 subscribers

10K views 2 years ago SAN FRANCISCO

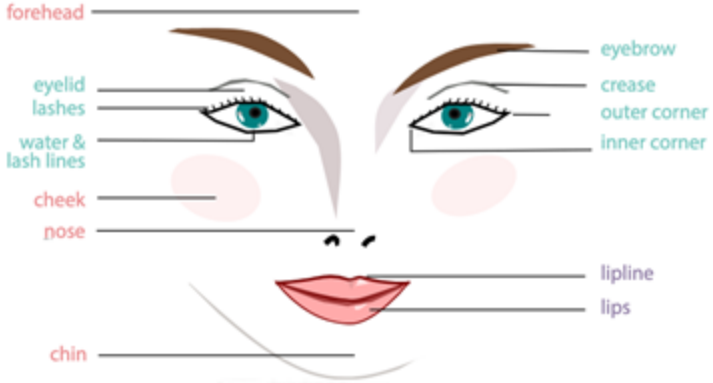
A seesaw is a staple game in most parks and playground. It is so fun, especially when you have somebody to share with.
Hayden and Irvin have started to play together much more, and so I thought it would be nice if they have a seesaw in the backyard, so they could hang out whenever they want.
In this 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

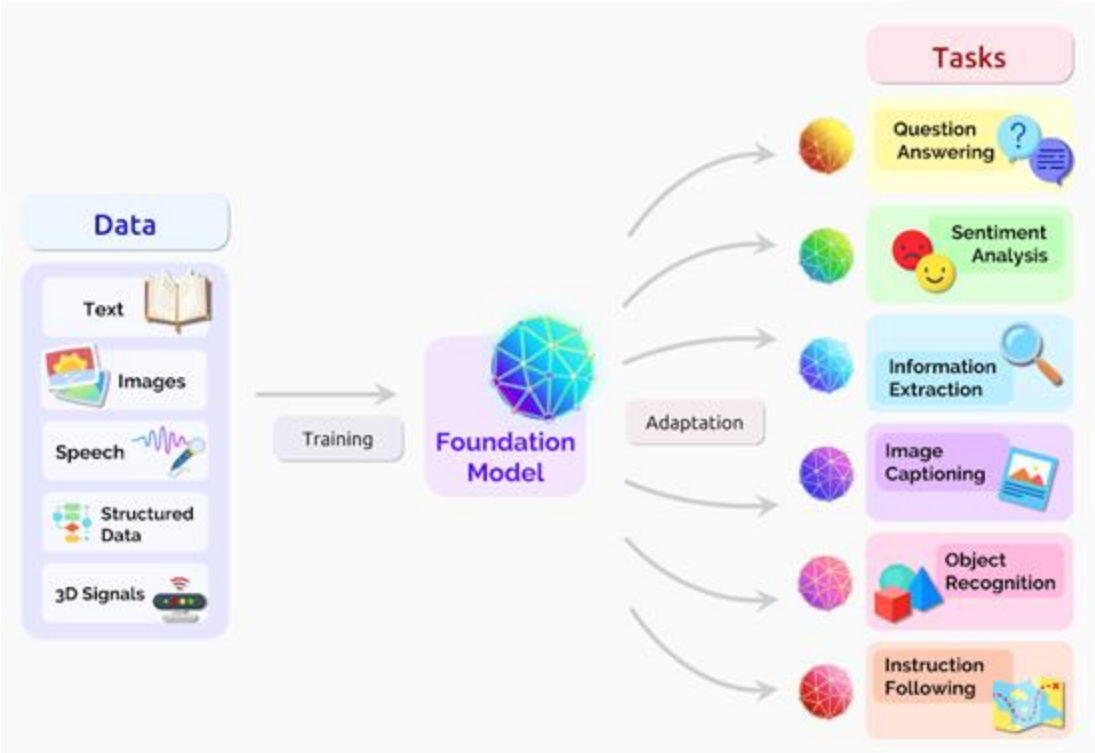
Creating high-quality tutorials require domain knowledge

For makeup tutorials,
cluster steps by facial parts



Truong et al. Automatic generation of two-level hierarchical tutorials from instructional makeup videos. CHI'21

AI presents opportunities to create cross-domain tutorials



Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." arXiv preprint 2021

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:

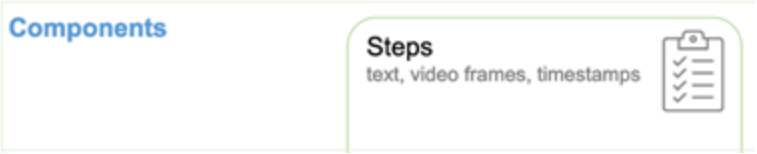
- Vast model space: multiple models are required and applicable
- Inaccurate results: models may make errors

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?

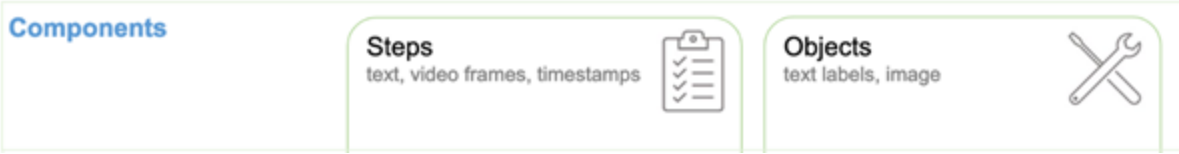
TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation

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

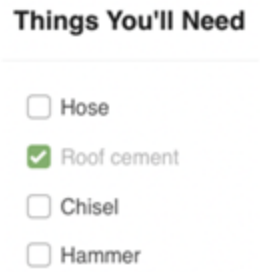


Examples:

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

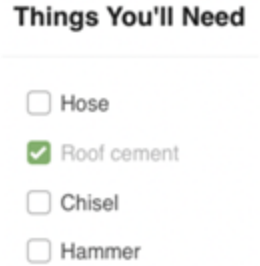


Examples:



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

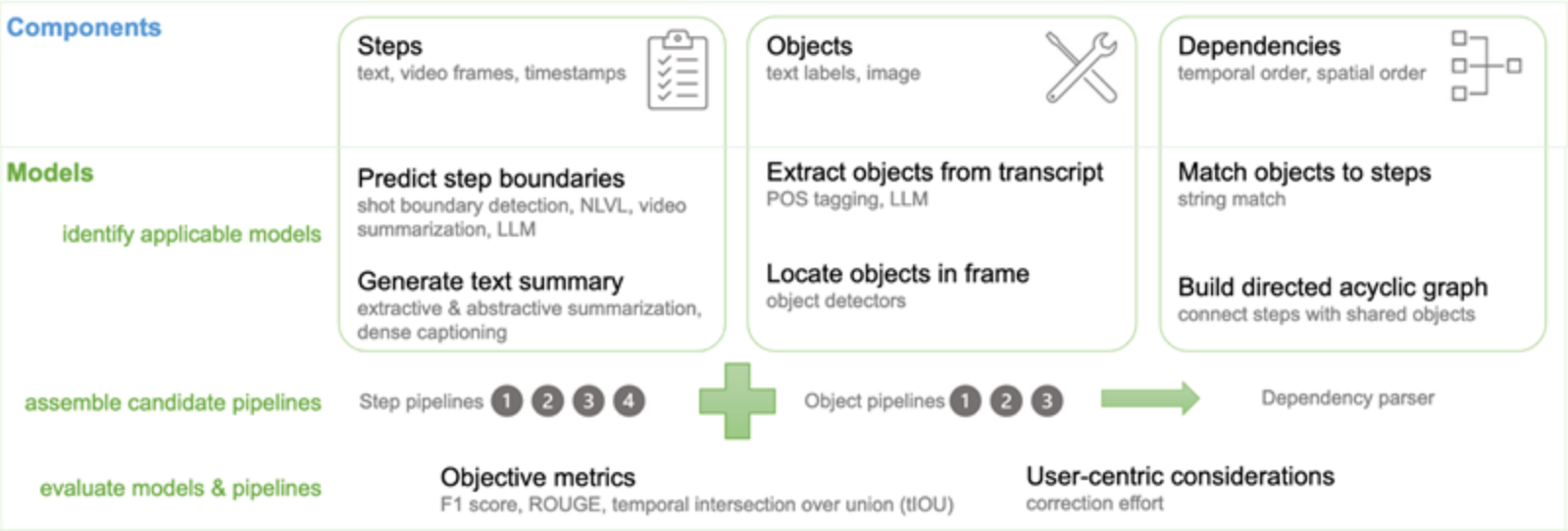
Examples:



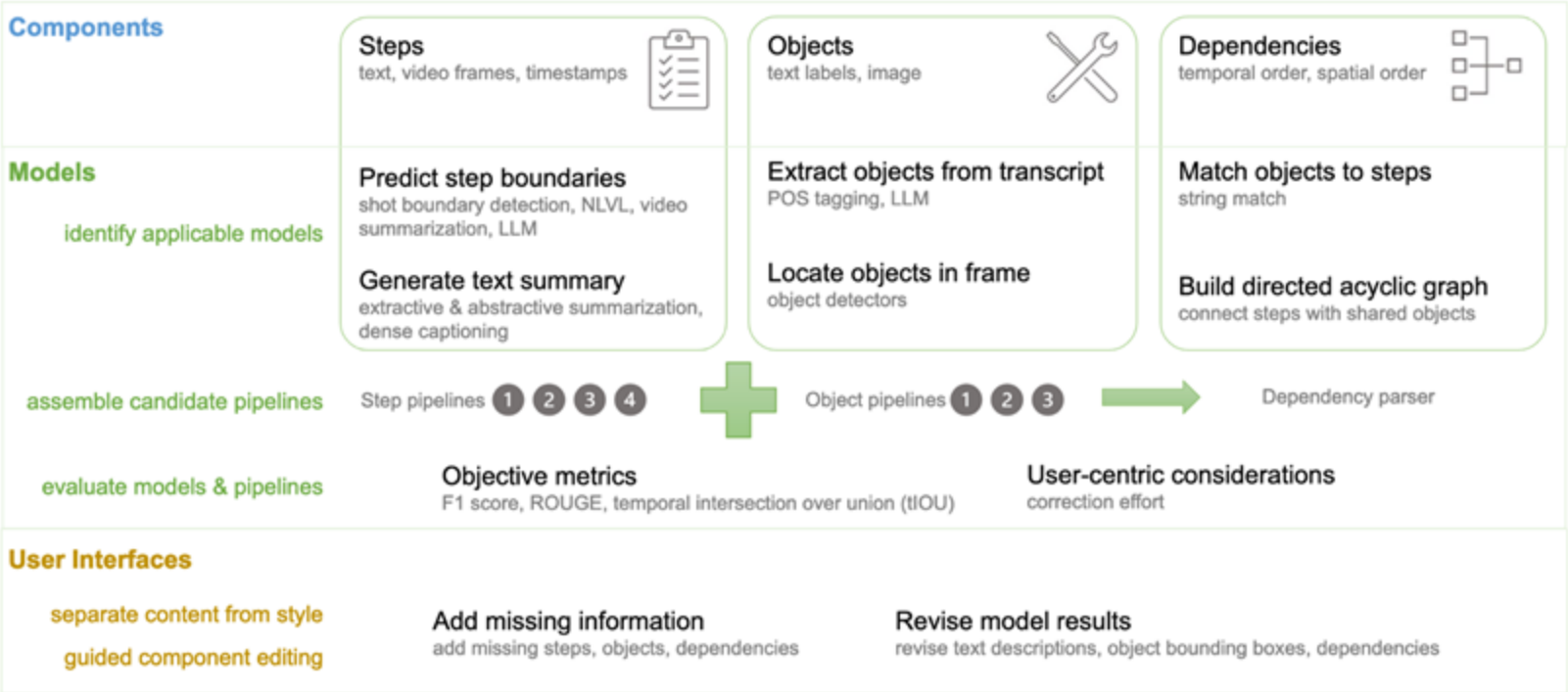
TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation

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

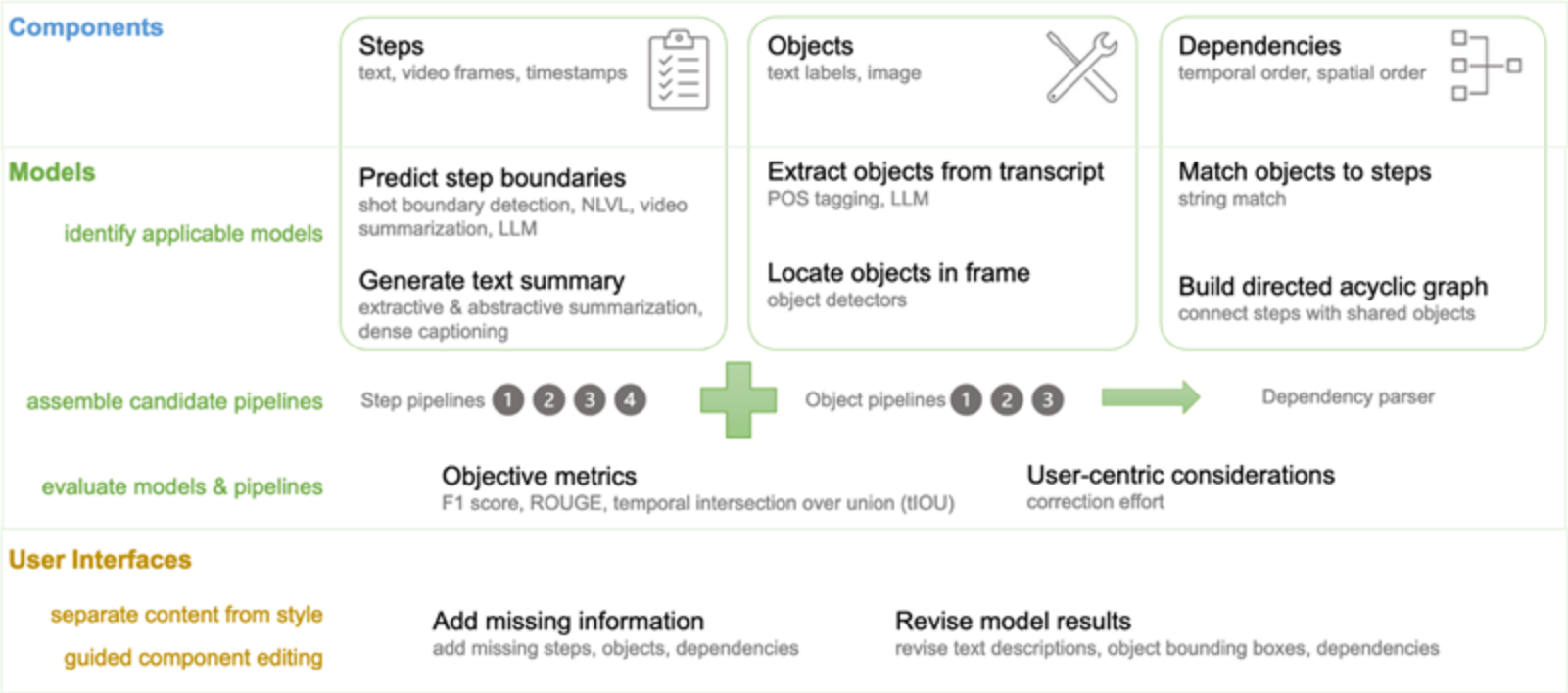
TutoAI: A cross-domain framework for AI-assisted mixed-media tutorial creation



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



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



Framework > Models > UIs > User Study > Guidelines > Future Work

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

Computation time on cpu: 1.228 s

dining table	0.978
bowl	0.998

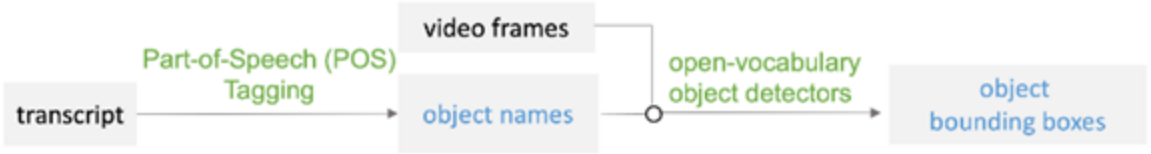
https://huggingface.co/facebook/detr-resnet-50

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

Candidate ML pipelines for object extraction

#1

#2



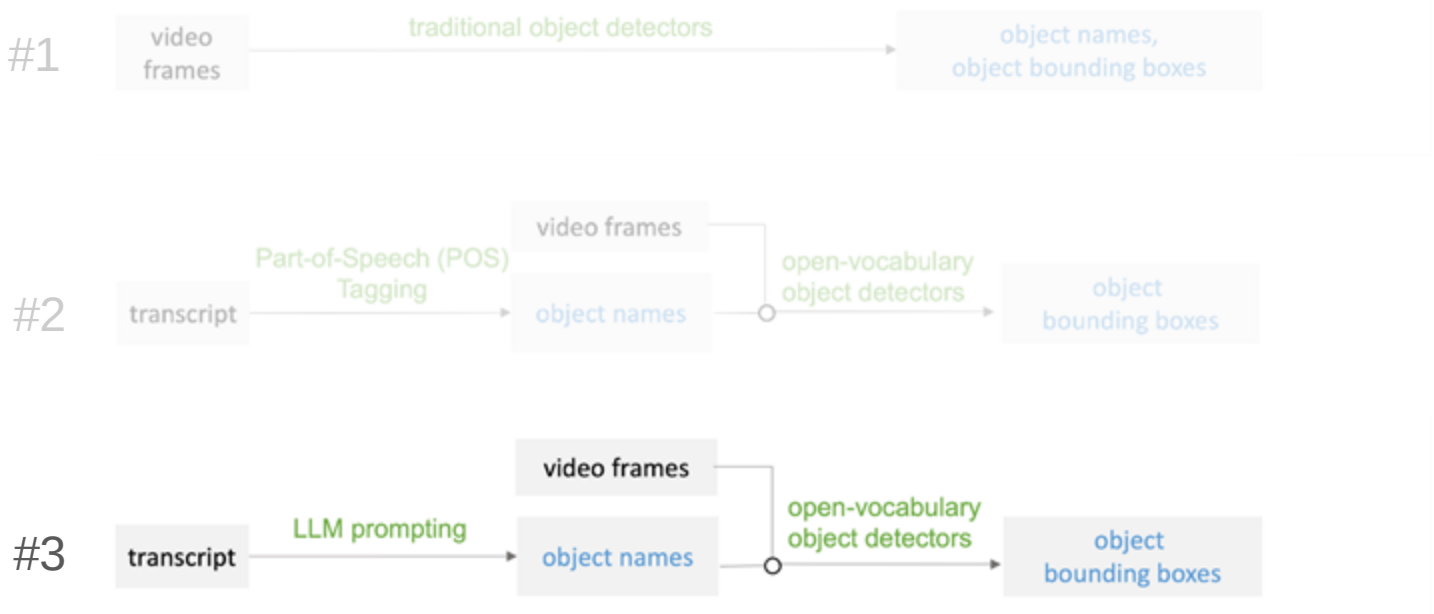
Pipeline #2: part-of-speech tagging

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

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

https://github.com/flairNLP/flair

Candidate ML pipelines for object extraction



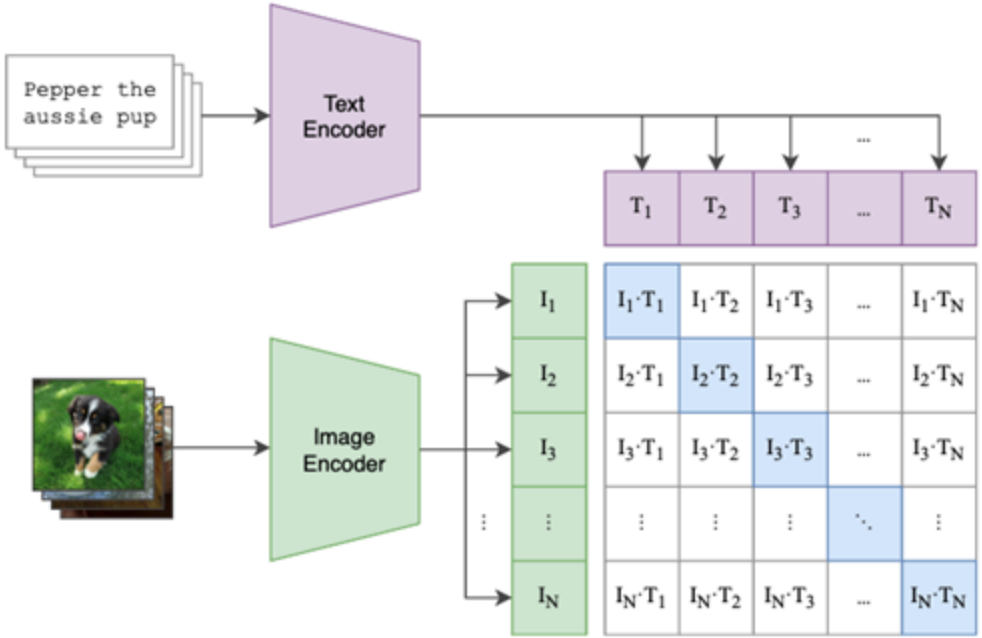
Pipeline #3: large language model (LLM) prompting

Input: the entire transcript

Prompt: “Identify the ingredients in this tutorial”

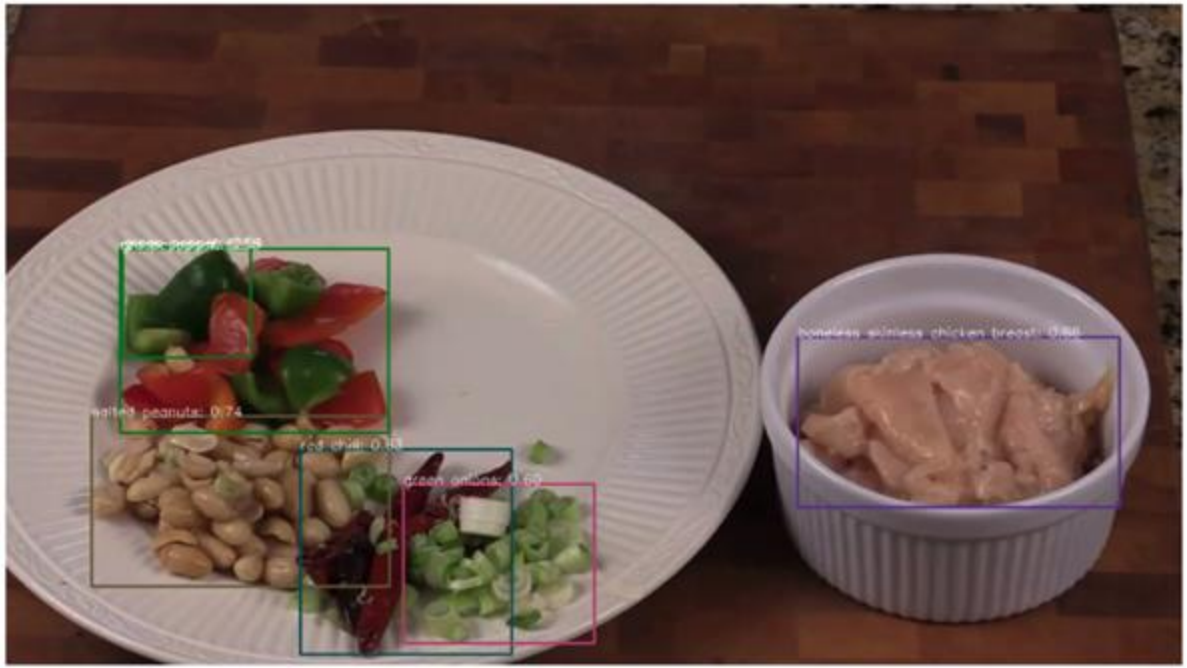
Output: “Boneless skinless chicken breast”, ...

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

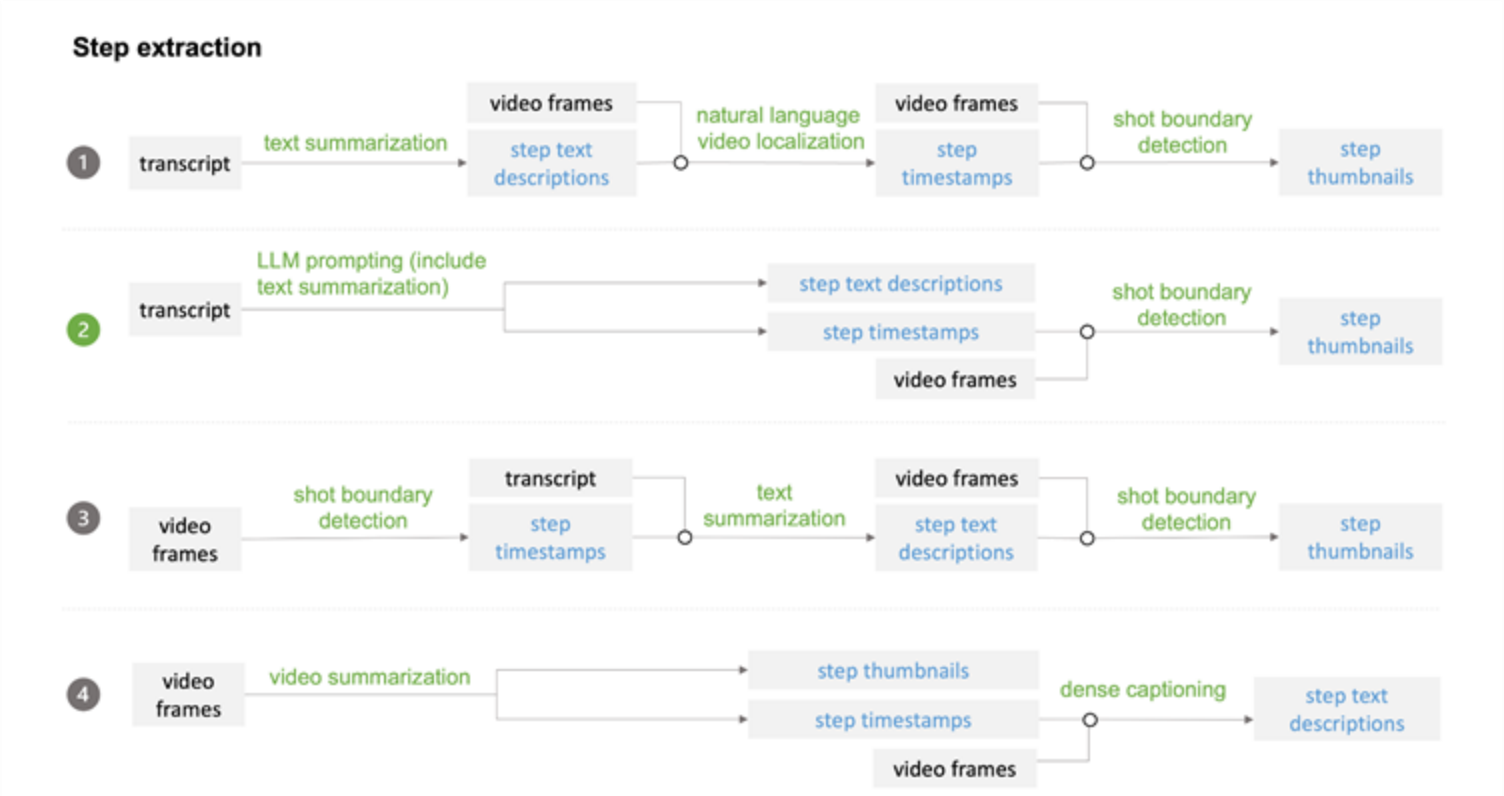


Labels extracted by LLM:

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

Li et al. "Grounded language-image pre-training." CVPR 2022

Candidate ML pipelines for step extraction



Evaluation: a large-scale cooking video dataset

Dataset:

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

YouCook2 Dataset

Evaluation: a large-scale cooking video dataset

Dataset:

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

YouCook2 Dataset

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

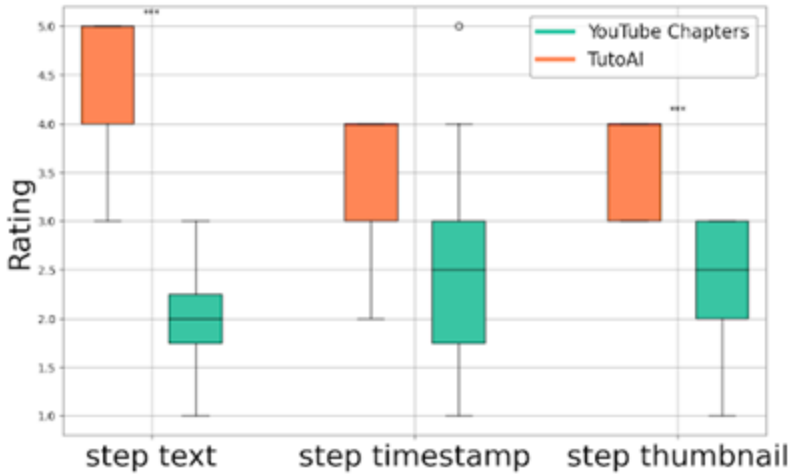
00:16 they have a seesaw in the backyard so they could hang out whenever they

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

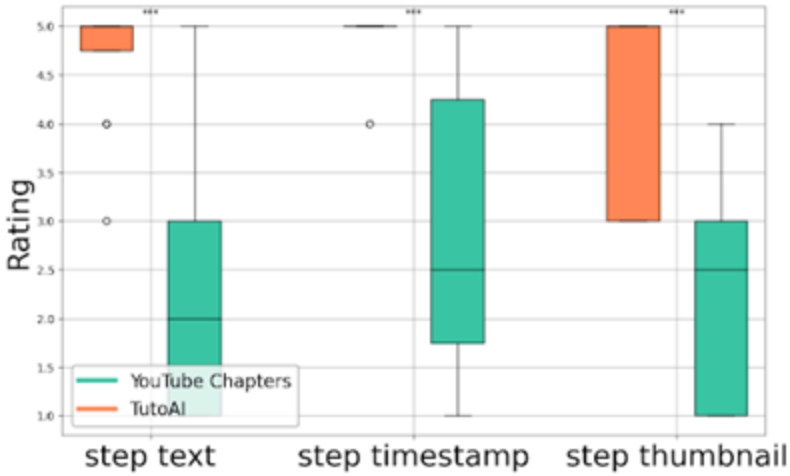
TutoAI vs. YouTube Chapters

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

Component quality **before editing**

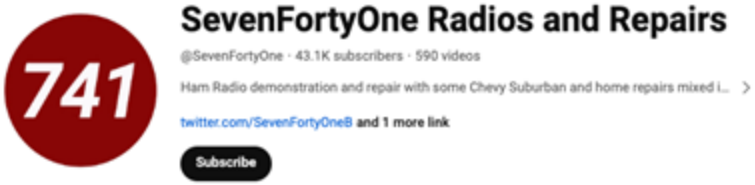


Component usefulness **after editing**



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…"*



TutoAI guidelines for human-centered system builders

TutoAI guidelines for human-centered system builders

Adopt a multi-modal perspective

INITIALLY

TutoAI guidelines for human-centered system builders

Adopt a multi-modal perspective

INITIALLY

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

INITIALLY

Focus on user-centric model selection

Support graceful degradation

WHEN WRONG

Simplify complex creation by guiding and constraining user actions

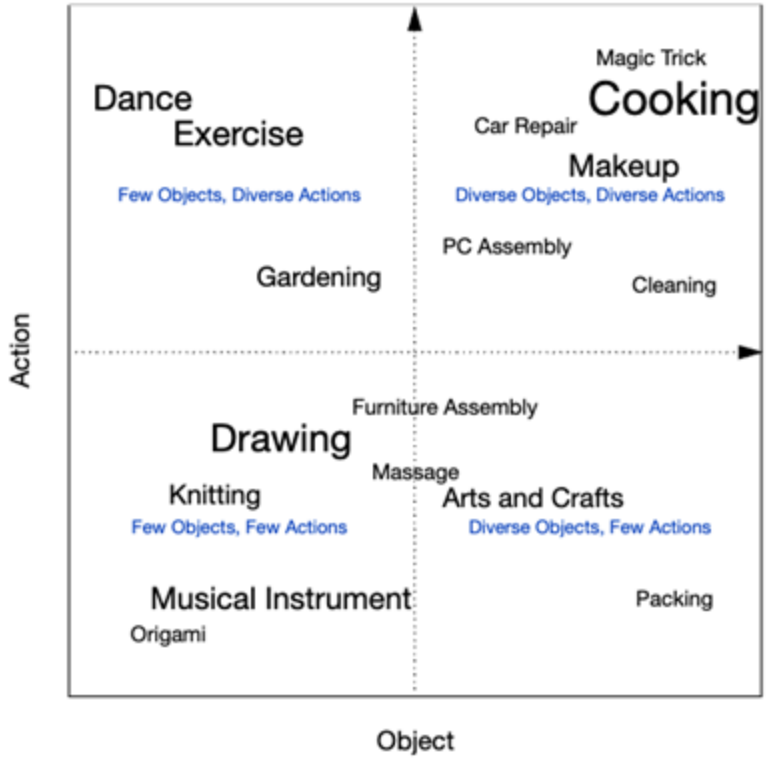
Separate content from styles

DURING INTERACTION

Leverage strong models for cross-modal enhancement

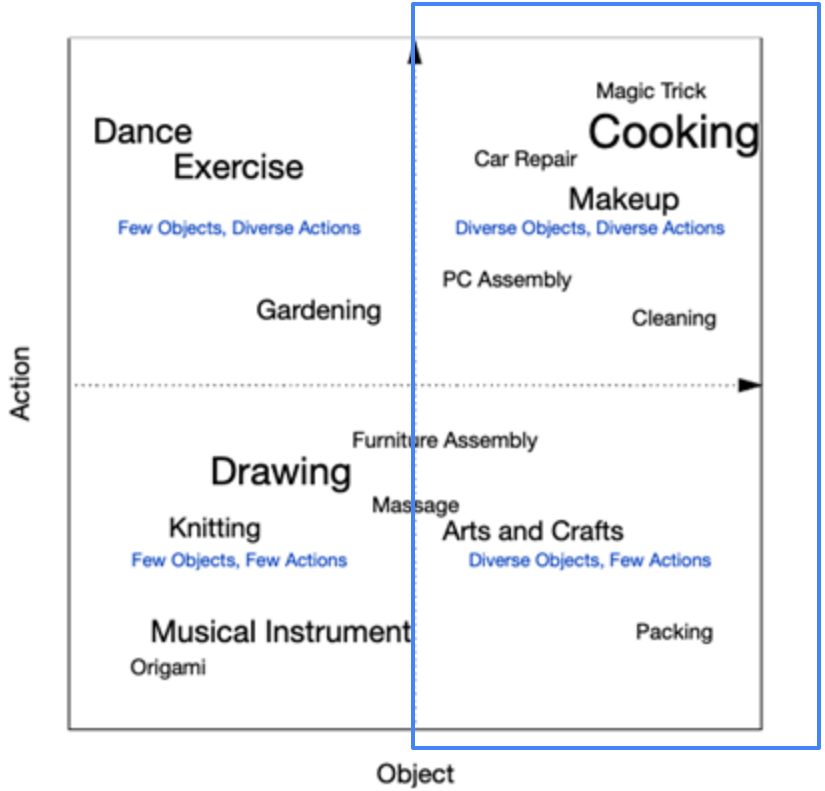
OVER TIME

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