# Investigate Procedure Events in *Multimodal* Fashion

**Presenter: Yue Yang**

*Department of Computer and Information Science, University of Pennsylvania*

Penn Engineering

# Introduction

- Procedural events: a set of steps to accomplish a certain goal.
- Represented as *scripts/schema* that human uses to perform everyday tasks.

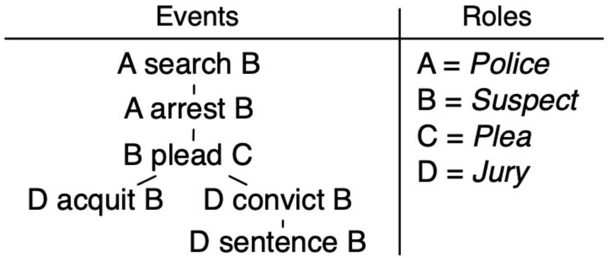
*Schema of Change a Tire*

- Find a safe place.
- Park the car.
- Take out the spare tire.
- Raise the jack.
- Loosen the nuts.
- ……

Flat tire during my trip in CA

# Introduction

- Scripts for natural language understanding (Schank and Abelson, 1977)
- Supervised learning for corpora, e.g., Framenet (Baker et al., 1998)
- Narrative Schemas and Event Chains (Chambers and Jurafsky, 2007, 2008, 2009)



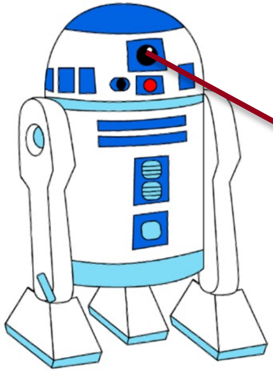

- Goal-Step Relations (Lyu et al., 2020)
  - Goal-Step Inference
  - Step Ordering

Goal: Prevent Coronavirus
**A. wash your hands** B. wash your cat
C. clap your hands D. eat your protein

Goal: Clean Silver
A. dry the silver
**B. handwash the silver**

# Motivation

- Past work mostly examined the procedure events for **text**.




Make Tea?
Make Coffee?
Cook Noodles?

Recommendations

# Motivation

Schema of *Get a slice of cake*:

take the cake out of the box → cut a slice → put it on a plate → take the plate to the user

Reporting Bias (Gordon and Van Durme, 2013)

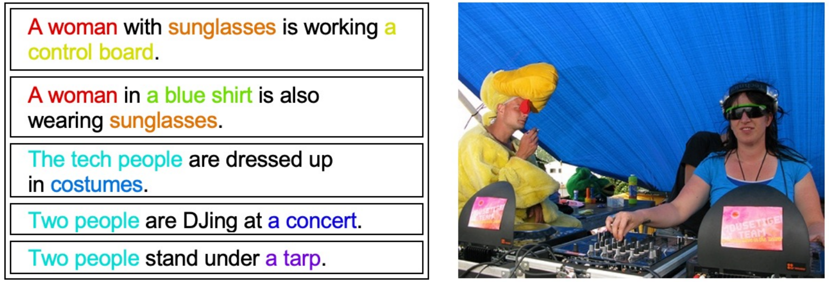
# Visual Goal-Step Inference using wikiHow

**Yue Yang**, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, Chris Callison-Burch

*Department of Computer and Information Science, University of Pennsylvania*

# Introduction

- Learning goal-step relations in multimodal fashion.
- We propose the Visual Goal-Step Inference (VGSI)
  - Given given a textual goal.
  - Infer which image represents a plausible step.
- More challenging than text-image matching
  - Text and objects are not closely matched




Caption-based image-text task

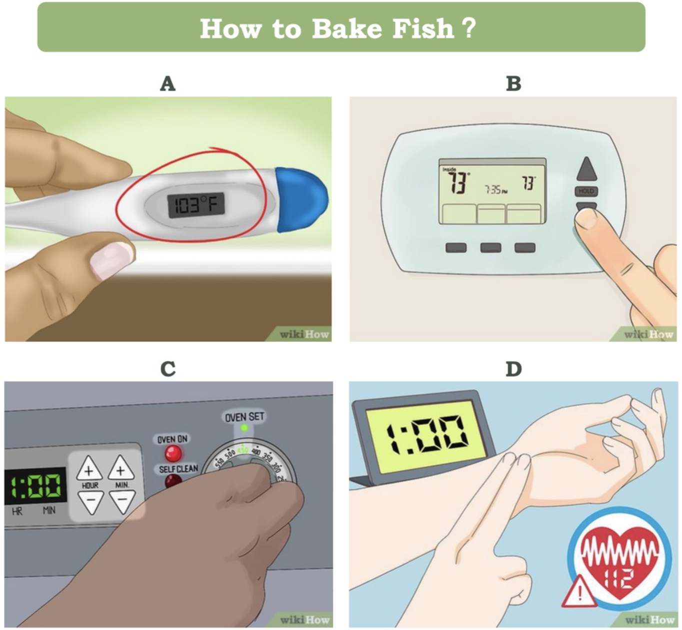



Figure 1: An example Visual Goal-Step Inference Task: given a text goal (*bake fish*), select the image (C) that represents a step towards that goal.

Penn Engineering

# Dataset

- Harvested from wikiHow
- Goal – Method – Step structure
- The corpus consists
  - 53,189 wikiHow articles across various categories
  - 155,265 methods, 772,294 steps/images

| Category | Goals | Methods | Steps | Images |
|---|---|---|---|---|
| Health | 7.8k | 19.1k | 97.5k | 111.8k |
| Home and Garden | 5.9k | 16.0k | 82.9k | 85.4k |
| Education & Communications | 4.7k | 12.4k | 61.2k | 66.1k |
| Food & Entertaining | 4.6k | 11.6k | 62.0k | 69.0k |
| Finance & Business | 4.4k | 11.8k | 59.3k | 66.8k |
| Pets & Animals | 3.5k | 9.5k | 45.3k | 48.0k |
| Personal Care & Style | 3.4k | 9.0k | 46.1k | 48.9k |
| Hobbies & Crafts | 2.8k | 7.5k | 40.9k | 42.7k |
| Computers & Electronics | 2.6k | 6.1k | 31.5k | 36.2k |
| Arts & Entertainment | 2.5k | 6.8k | 35.4k | 37.2k |
| Total | 53.2k | 155.3k | 772.3k | 772.3k |

Table 1: Number of goals, methods, steps and images in the top 10 wikiHow categories.

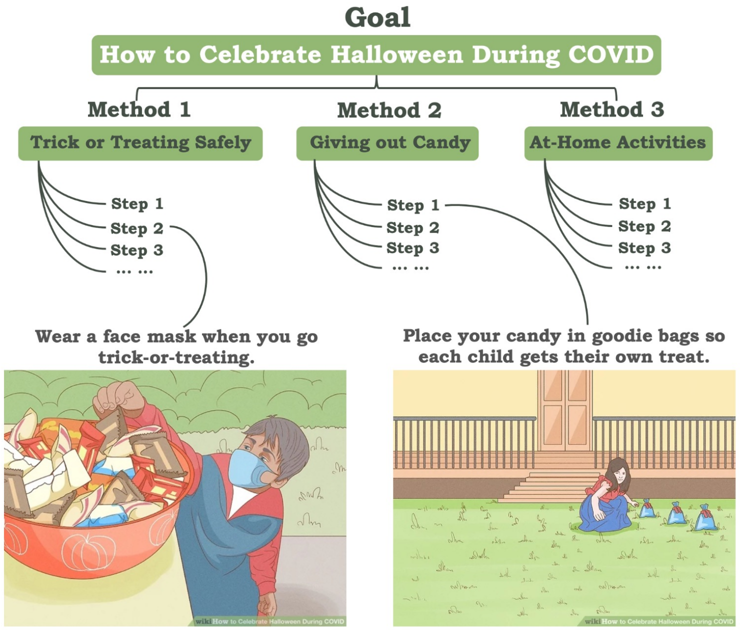
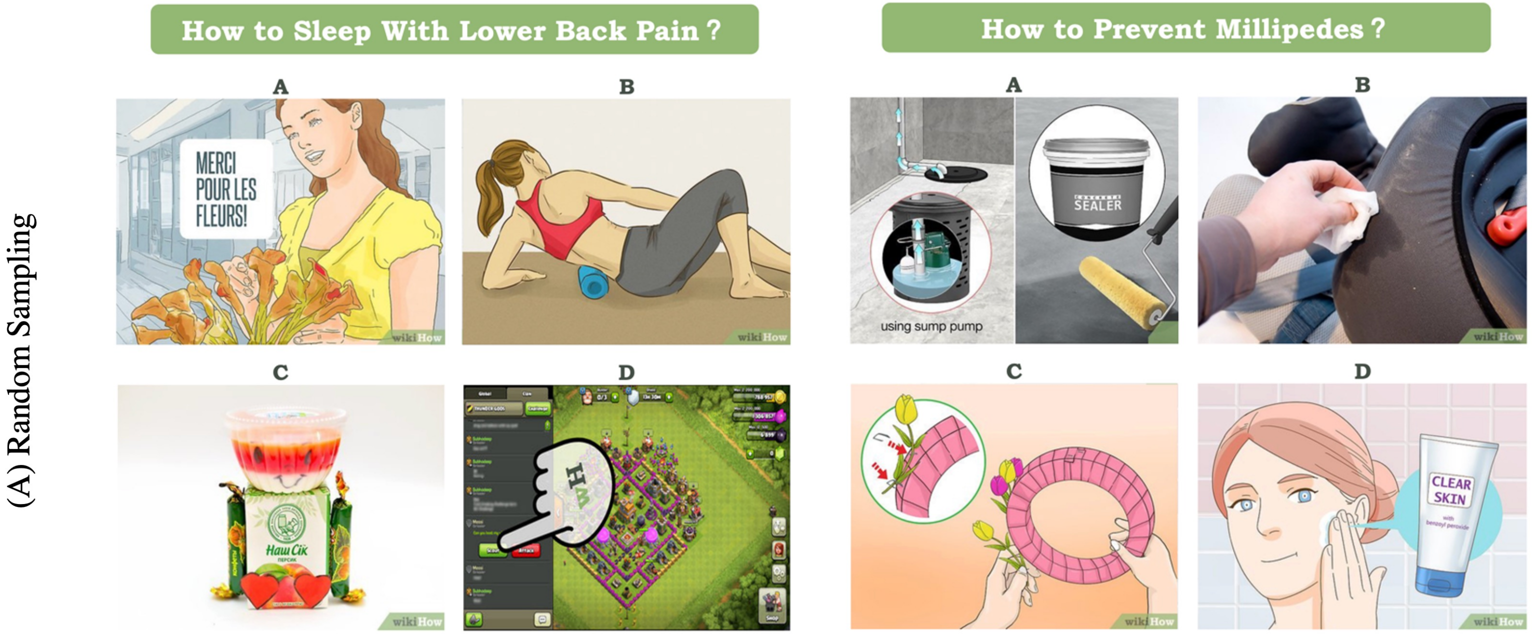



Figure 2: Hierarchical multimodality of wikiHow.

# Sampling Strategies

- Random Sampling



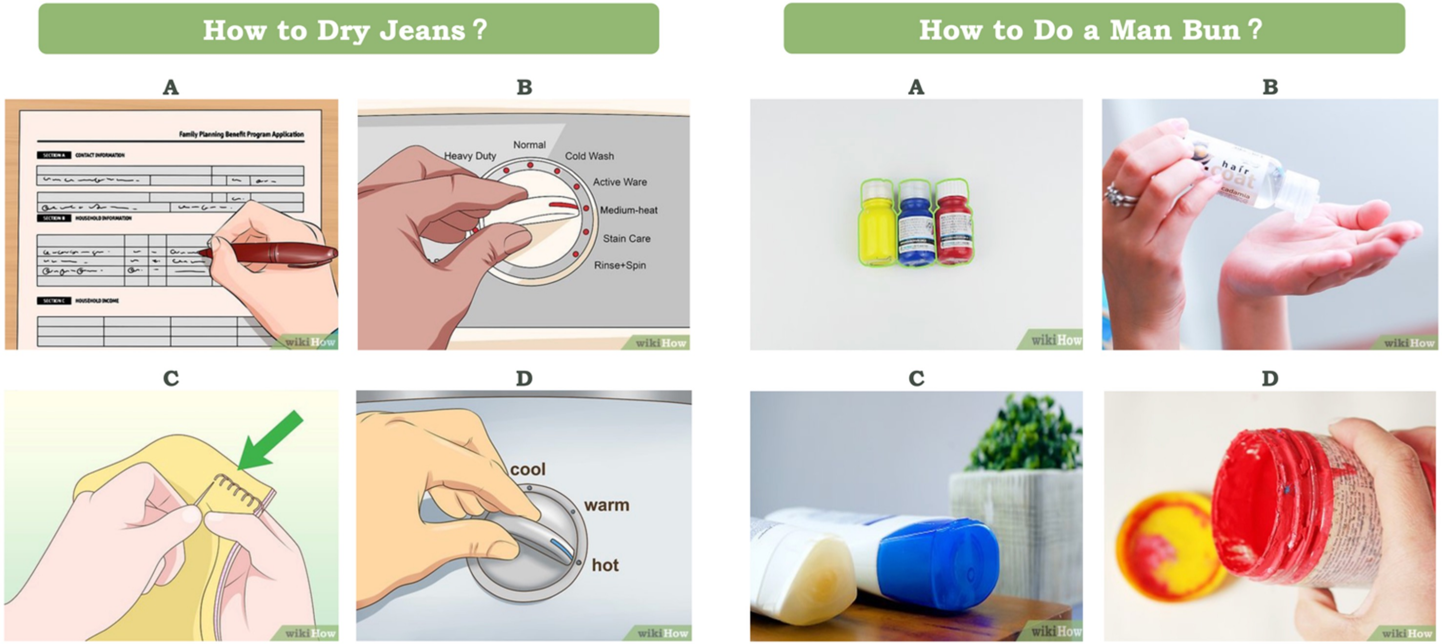

(A.1) Correct Answer is **B**

(A.2) Correct Answer is **A**

# Sampling Strategies

- Similarity Sampling

(B) Similarity Sampling



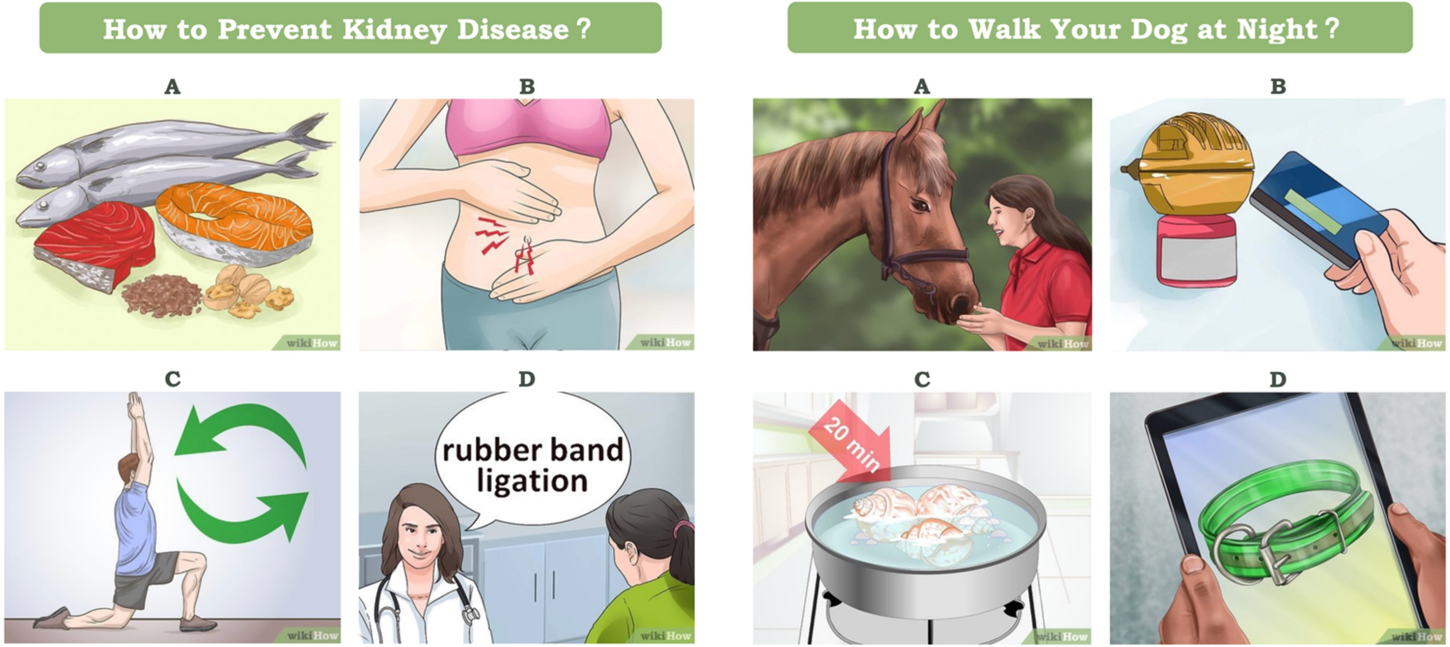

(B.1) Correct Answer is **B**

(B.2) Correct Answer is **C**

# Sampling Strategies

- Category Sampling

(C) Category Sampling




(C.1) Correct Answer is **A**

(C.2) Correct Answer is **D**

# Experiments

- Problem Formulation:
  - Input: a high-level goal $G$, an Image $I$
  - The model outputs the matching score:

$$match(G, I) = F(X_G, X_I) \qquad (1)$$

- Baseline Models:
  - **DeViSE**
  - **Similarity Network**
  - **Triplet Network**
  - **LXMERT (transformer-based)**

- Human Annotation

# Results

| Model | Sampling Strategy (Test Size) | | |
|---|---|---|---|
| | Random (153,961) | Similarity (153,770) | Category (153,961) |
| Random | .2500 | .2500 | .2500 |
| DeViSE | .6719 | .3364 | .4558 |
| Similarity Net | .6895 | .6226 | .4983 |
| LXMERT | .7175 | $\boxed{.4259}$ | $\boxed{.2886}$ |
| Triplet Net (GloVe) | .7251 | .7450 | .5307 |
| Triplet Net (BERT) | **.7280** -13.8% | **.7494** -8.77% | **.5360** -29.0% |
| Human | .8450 | .8214 | .7550 |

Table 2: Accuracy of SOTA models on the wikiHow VGSI test set with different sampling strategies (sample size is shown in parentheses).

# Results

- The knowledge learned from wikiHow can be transferred to other datasets.

| | | Sampling Strategy | | |
|---|---|---|---|---|
| **PT-Data** | **FT?** | **Random** | **Similarity** | **Category** |
| - | ✓ | .6005 | .6096 | .4434 |
| Flickr30K | ✗ | .4837 | .5398 | .3856 |
| | ✓ | .6207 | .6408 | .4740 |
| MSCOCO | ✗ | .5099 | .5715 | .3958 |
| | ✓ | .6340 | .6640 | .4794 |
| COIN | ✗ | .5067 | .5161 | .3978 |
| | ✓ | .6170 | .6343 | .4638 |
| wikiHow | ✗ | **.6556** | **.6754** | **.4750** |
| | ✓ | **.6855** | **.7249** | **.5143** |
| Human | - | .8300 | .7858 | .7550 |

Table 4: Transfer performance (4-way multiple choice accuracy) on Howto100m. FT results are obtained by fine-tuning the model on the full training set.

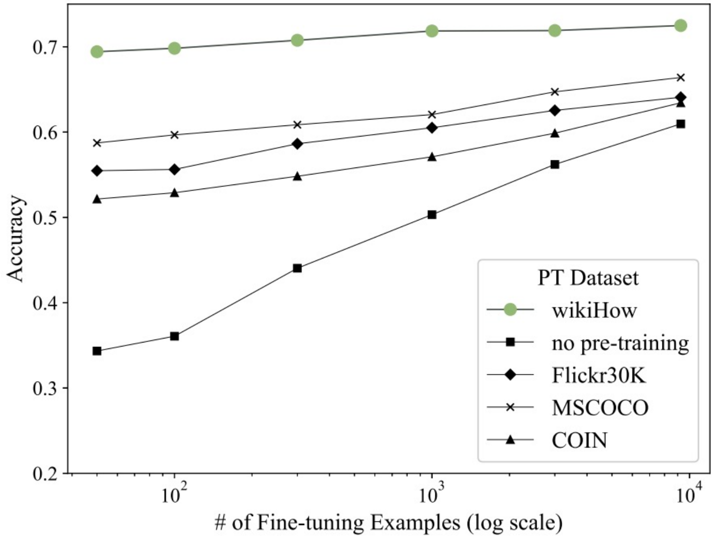



Figure 5: Transfer performance on Howto100m (similarity sampling) with different pre-training datasets vs. the number of training examples.

Penn Engineering

# Conclusion

- We propose the novel Visual Goal-Step Inference task (VGSI), a multimodal challenge for reasoning over procedural events.
- We construct a dataset from wikiHow and show that SOTA models struggle on it.
- The knowledge harvested from our dataset could be transferred to other datasets.
- The multimodal representation learned from VGSI has strong potential to be useful for NLP applications such as multimodal dialog systems, multimodal schema induction systems.

Dataset and code are available at
https://github.com/YueYANG1996/wikiHow-VGSI

Penn Engineering

# Induce, Edit, Retrieve:
# Language Grounded Multimodal Schema for Instructional Video Retrieval

**Yue Yang**, Joongwon Kim, Artemis Panagopoulou,
Mark Yatskar, Chris Callison-Burch
*Department of Computer and Information Science, University of Pennsylvania*

Penn Engineering

# Motivation

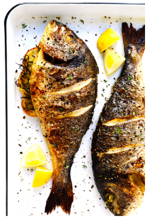
- *Schema*: a set of rules people use to perform everyday tasks.
- Schema can be *generalized.*
- When facing new tasks, people use *prior* knowledge. (Chen et al., 2004)

Bake *Cheesecake*

Bake *Fish*?

Bake *Cake*

Bake *Cupcake*



Bake *Cookies*

# Motivation

- *Problems on current Schema Induction System*
  - Use text only (reporting bias)
  - Rely on labeled data
  - Small scale
  - Multimodal downstream tasks

- *Can Vision adopt such reasoning approach?*
  - Induce schemata from visual signals.
  - Generalize schemata for larger scale.
  - Use schemata to improve multimodal tasks.

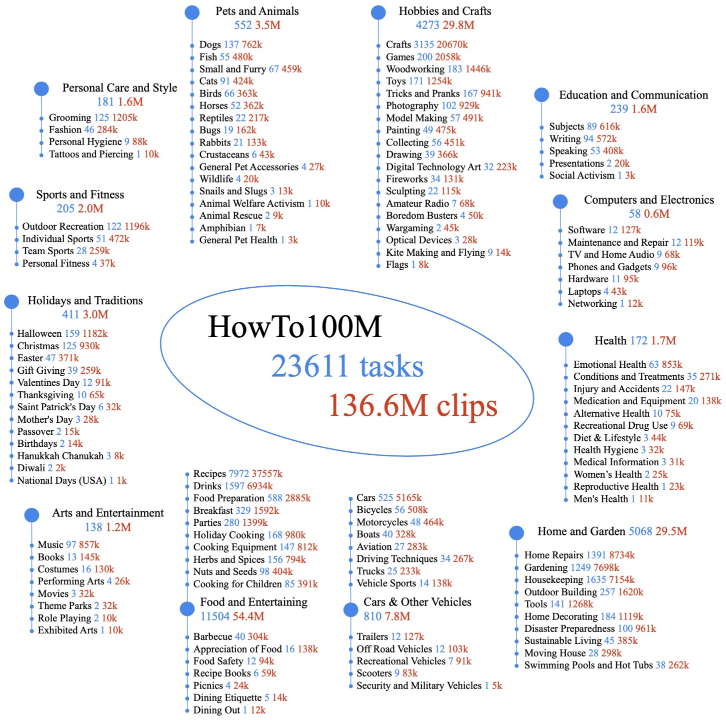
# IER Overview

- Our **I**nduce, **E**dit, **R**etrieve (IER) system:
  - Induce:
    - Input: A task name, a set of related videos
    - Output: A set of sentences as the schema
  - Edit
    - Given an unseen task
    - Use language models to modify schema
  - Retrieve
    - Improve video retrieval using schema


Figure 1. An example from our IER system, which first induces a schema for *Bake Chicken* using a set of videos. Then it edits the steps in the schema to adapt to the unseen task *Bake Fish* (the tokens that have been edited are highlighted). Finally, IER relies on the edited schema to help retrieve videos for *Bake Fish*.

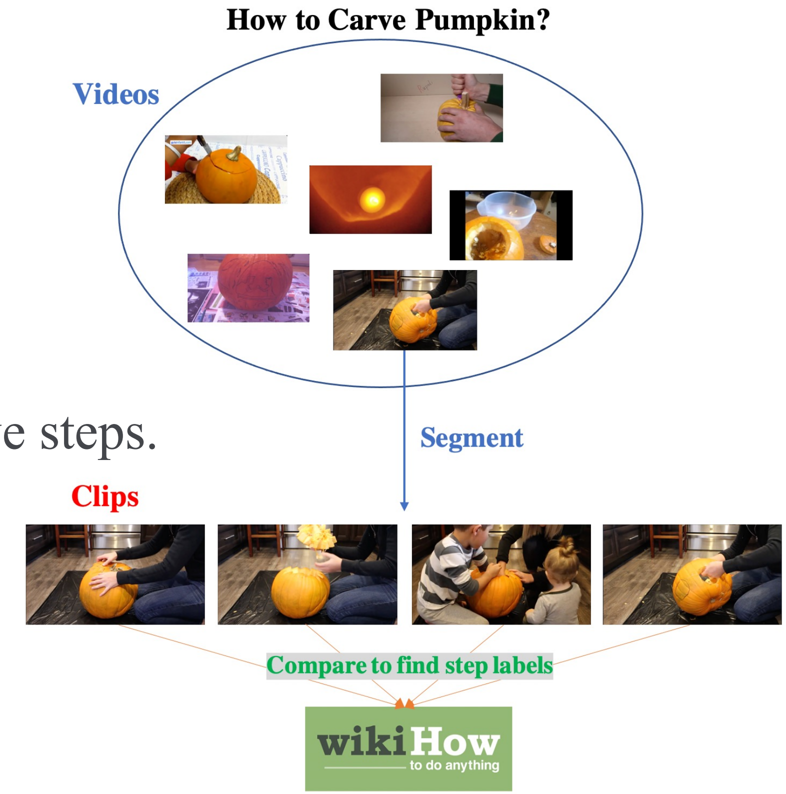
# Schema Induction

- Generate schemata for a set of tasks based on their associated videos
  - Input: A bag of videos
  - Out: A bag of step descriptions
- Learning Data – Howto100M (Miech et al., 2019)
  - 136 M video clips from 1.22M instructional videos
  - 23K tasks, directly from wikiHow
  - Focus on visual tasks only
  - Retrieve videos from YouTube



# Schema Induction

- How to covert video to text?
  - Captioning? Transcripts? Template?
- Use human written steps in wikiHow!
  - 1M steps from 110k articles on everyday tasks
- Use pretrained video-text model to retrieve steps.
  - MIL-NCE (Miech et al., 2019)
  - (Clip, Step) matching score
  - Pair every video segment with all 1M steps
  - Sort steps based on the matching score



# Schema Induction

**Task**: Make Tea
**Retrieved step label**: Strain the tea through a filter and pour it into cups.

**Task**: Carve Pumpkin
**Retrieved step label**: Scoop the seeds out of your pumpkin with a large serving spoon.

**Task**: Change a Tire
**Retrieved step label**: Jack the car up so that you can fit, comfortably, underneath the car.

# Schema Induction

- For each video segment, we select top-30 steps.
- We further sort these steps based on the average matching score across all videos.
- The top-100 steps are selected.
- Hierarchical clustering to remove paraphrases
- On average, 25.1 steps per task

Clip

Pair with steps

wikiHow to do anything

Pre-trained model

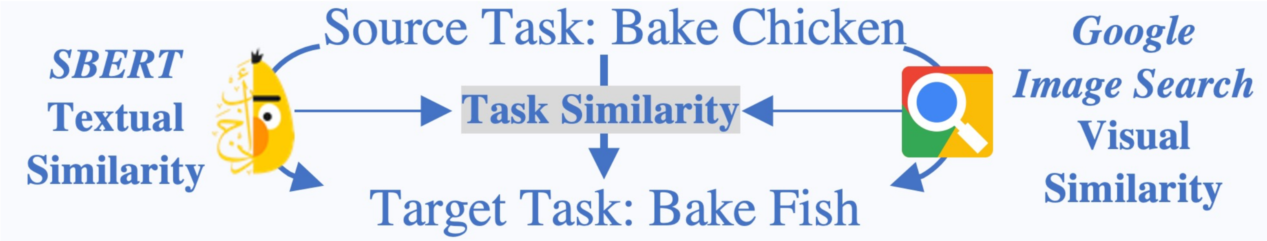
| Top-10 Retrieved Steps | Similarity Score |
|---|---|
| **Carve** a **scary pumpkin** and place it outside. | 13.25 |
| **Mark** the pumpkin **drill holes**. | 12.15 |
| **Carve a face** in the jelly, similar to a pumpkin face. | 11.93 |
| **Cut a hole** in the back of the pumpkin large to fit your hand through. | 11.83 |
| Carve a pumpkin for **Halloween**. | 11.50 |
| Carve **around the nose** with a **knife** to finish outlining it. | 11.36 |
| Place the **Cylon pumpkin** on display. | 11.35 |
| **Carve** the **pumpkin** as shown in this image. | 10.91 |
| Have some pumpkin **flesh and seeds**! | 10.29 |
| **Cut** the pumpkin in **half**, lengthwise. | 10.08 |

# Schema Induction

| How to Stain Cabinets | How to Use a Drill Safely | How to Replace Shocks |
|---|---|---|
| • **Glaze** the doors using the same process you did with the cabinets.<br>• Choose a whitewash **wood stain**.<br>• **Paint** dated **cabinets** and dark walls.<br>• Finish the **cabinets** with a **top coat**.<br>• Apply **glaze** to a section of one **cabinet door** or drawer.<br>• Opt for semi-custom cabinets for a midrange budget option with more features.<br>• **Prime the cabinets** with white primer paint.<br>• Put a lazy susan in your cabinets.<br>• Choose an appropriate **urethane** finish for the door.<br>• Apply the dye to the poplar with a **rag**. | • Set the **plunge depth** for the drill.<br>• Put on **safety glasses** before you start drilling.<br>• **Secure the cord grip** by installing the grub screw with an **Allen wrench**.<br>• Wear **safety goggles** and **a dust mask** while drilling.<br>• **Locate** the **chuck** at the end of the drill.<br>• Drill your team with simulated data breaches.<br>• Drill through the tile slowly.<br>• Set up your guide rail for cutting with a plunge saw.<br>• Complete **routing** and other machining before ebonizing.<br>• Wear the proper **safety gear** when **sawing and drilling** into wood. | • Visually **inspect** your **strut mounts** or **shock towers**.<br>• Call the bank's toll-free customer service number.<br>• Sign up for an email service.<br>• Drop it off at an **auto repair** or **auto parts shop**.<br>• Replace each hubcap.<br>• Inspect your **wheel** wells and **bumpers**.<br>• Examine the lug nuts.<br>• Take your vehicle to a **reputable repair shop** for **diagnosis and repairs**.<br>• Keep your tires aired up.<br>• Loosen the bleeder. |

# Schema Editing

- Given an unseen task without videos, edit existing schema.
- Find the most similar task in the schema library




- Textual Similarity = cosine similarity of SBERT embeddings
- Visual Similarity (Google Image Search)
- Task Similarity = max(Textual Similarity, Visual Similarity)

# Schema Editing – Object Replacement

- Editing Module 1: **Object Replacement**
- Every task has a main object, e.g., "chicken" of "Bake Chicken"
- Use POS tagger to find the 1st occurred noun as main object
- Replace the objects in all steps

| **Object Replacement** |
| --- |
| Cook Ham $\xrightarrow{0.86}$ Cook Lamb<br>Put the ham in the oven.<br>↓<br>Put the lamb in the oven. |
| Clean a Guitar $\xrightarrow{0.84}$ Build a Violin<br>Use a polish for particularly dirty guitars.<br>↓<br>Use a polish for particularly dirty violins. |
| Trap a Rat $\xrightarrow{0.84}$ Trap a Rabbit<br>Bait and set snap rat traps.<br>↓<br>Bait and set snap rabbit traps. |

# Schema Editing – Step Deletion

- Editing Module 2: **Step Deletion**
- Delete the steps no longer suitable for the new target task.
- "Insert a roasting thermometer into the thigh" of "Bake Chicken" ❌ "Bake Fish"
- Sentence BERT pretrained on question-answer pairs.
- Compute the score of (task, step).
- if (source task, step) >> (target task, step)
  delete, otherwise include

| **Step Deletion** |
|---|
| Transplant a Young Tree $\xrightarrow{0.89}$ Remove a Tree |
| Fill your pot with a balanced fertilizer. |
| $\downarrow$delete |
| ~~Fill your pot with a balanced fertilizer.~~ |
| Fix a Toilet $\xrightarrow{0.85}$ Remove a Toilet |
| Test out the new flapper. |
| $\downarrow$delete |
| ~~Test out the new flapper.~~ |
| Brush a Cat $\xrightarrow{0.87}$ Brush a Long Haired Dog |
| Comb and groom your pet. |
| $\downarrow$include |
| Comb and groom your pet. |

# Schema Editing – Token Replacement

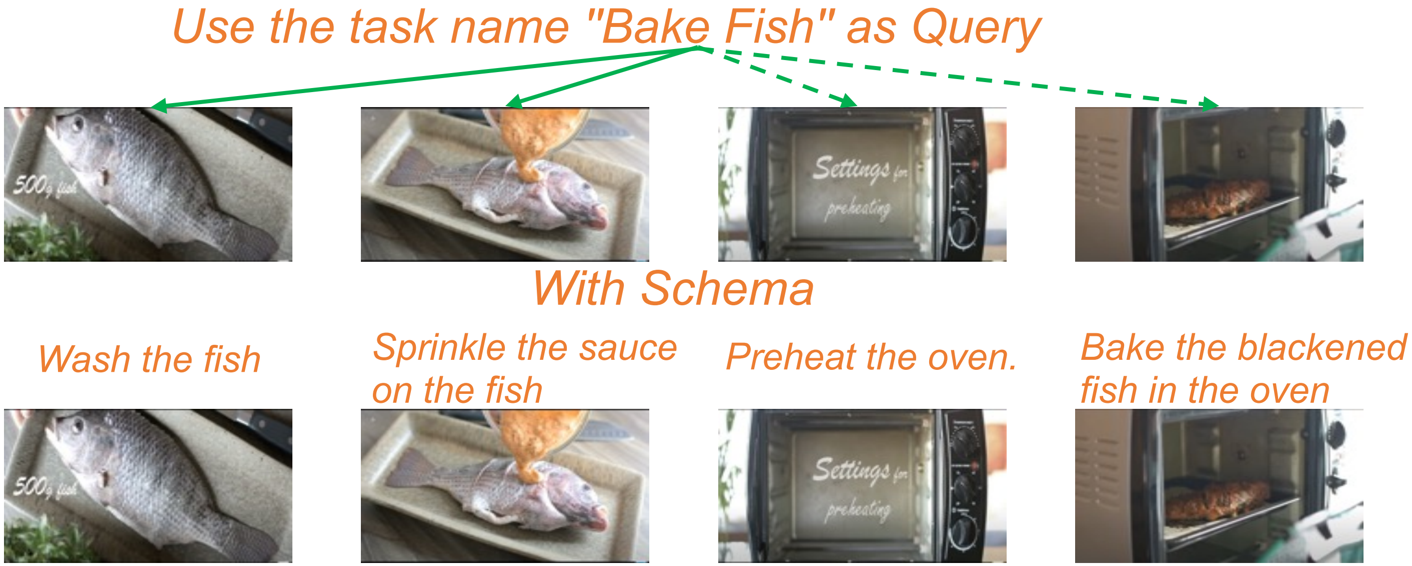
- Editing Module 3: **Token Replacement**
- Use *masked language model* to replace the token with the lowest probability.
  - "Season the drumstick" in "Bake Chicken"
  - Mask the token "Season the <mask>".
- Use a prompt: How to [TASK]? [STEP]
  - How to Bake Fish? "Season the <mask>".
- Predict a new token from vocabulary
  - <mask> → fish, "Season the fish"

| **Token Replacement** |
|---|
| Prepare Fish $\xrightarrow{0.82}$ Prepare Crabs<br>Cut the fins from the fish using kitchen shears.<br>↓<br>Cut the shells from the crabs using steel scissors. |
| Make Healthy Donuts $\xrightarrow{0.88}$ Bake Healthy Cookies<br>Slice your donuts into disks.<br>↓<br>Slice your cookies into squares. |
| Wash Your Bike $\xrightarrow{0.84}$ Wash a Motorcycle<br>Clean the bike chain with a degreaser.<br>↓<br>Clean the motorcycle thoroughly with a towel. |

Penn Engineering

# Schema Guided Video Retrieval

- Query: Task Name (short)  Retrieve **long multi-minute** videos
- Global Matching (use task name only)
- Step Aggregation (use schemata to expand task name)



# IER Review

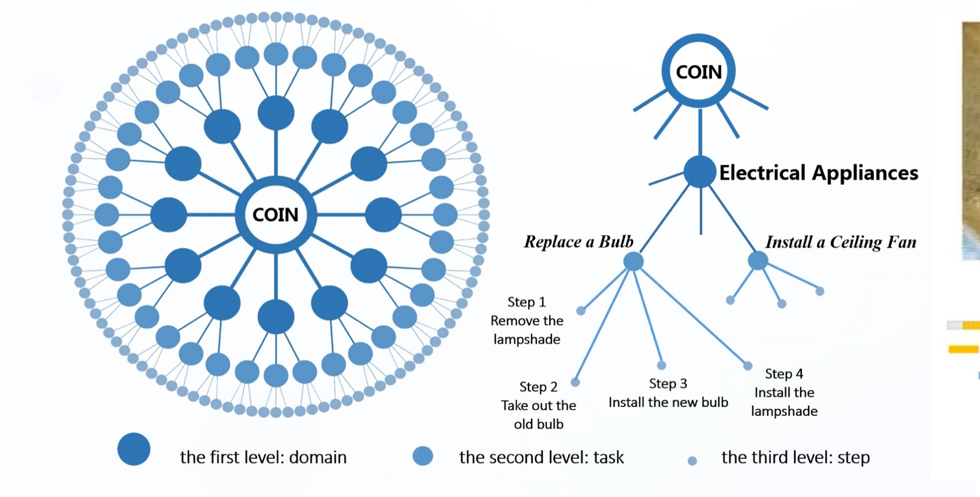
Schema Induction
Objective: Construct schemata on tasks.
Task: Bake Chicken
YouTube Videos (Learning Data)
Video 1
Video n
Clip-Step Alignment
wikiHow 1M human written steps
Wash the chicken thoroughly.
Marinate the chicken Season the drumsticks
Bake the blackened chicken in the oven or on the grill
Schema of Bake Chicken
Wash the chicken thoroughly
Season your drumsticks
Marinate the chicken
Insert a roasting thermometer into the thigh
Sprinkle the ginger on the chicken
Bake the blackened chicken in the oven
... ...
Schema Editing
Objective: Edit existing schema for unseen task.
Source Task: Bake Chicken
SBERT Textual Similarity
Task Similarity
Google Image Search Visual Similarity
Target Task: Bake Fish
Object Replacement
Source Step: Wash the chicken thoroughly
Target Step: Wash the fish thoroughly
Step Deletion
Source Step 1: Insert a roasting thermometer into the thigh
P(Source Step 1 | Bake Fish) << P(Source Step 1 | Bake Chicken)
Delete this step.
Source Step 2: Bake the blackened chicken/fish in the oven
P(Source Step 2 | Bake Fish) ≈ P(Source Step 2 | Bake Chicken)
Include this step.
Token Replacement
Source Step: Sprinkle the ginger on the fish
Mask the token: Sprinkle the <mask> on the fish
Use LM predict a new token:
Target Step: Sprinkle the sauce on the fish
Predicted Schema of Bake Fish
Wash the fish thoroughly
Season your fish
Marinate the fish
Sprinkle the sauce on the fish
Bake the blackened fish in the oven
... ...
Schema-Guided Video Retrieval
Objective: Use schema to improve retrieval performance.
Retrieve videos of "Bake Fish".
Use the task name "Bake Fish" as query.
... ...
Global Matching: Lack of Intermediate Information.
With Schema:
Predicted Schema of Bake Fish
Wash the fish thoroughly
Season your fish
Marinate the fish
Sprinkle the sauce on the fish
Bake the blackened fish in the oven
... ...
Wash the fish thoroughly
Sprinkle the sauce on the fish
Bake the blackened fish in the oven
... ...

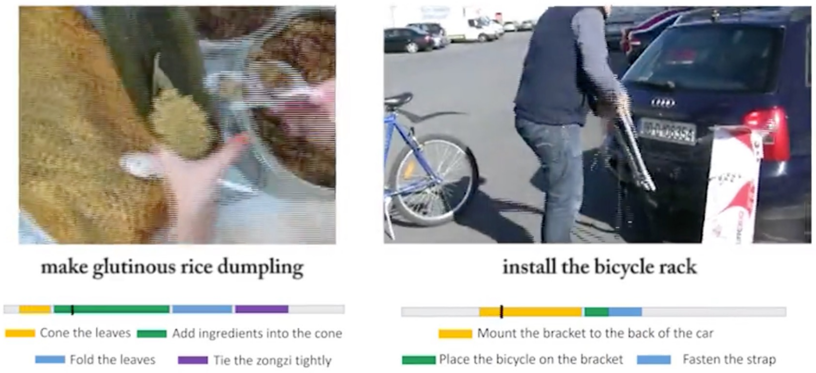
# Experiments - Datasets

- **Howto-GEN** (a new split of Howto100M)
  - Select the task names with exact one noun
    - 3365 tasks, 2,184 unique main objects
  - Random select 500 tasks for training, 500 for validation, 2365 for test.
  - 1,088 unseen main objects in the test set
    - Train: Peel Tomato
    - Test: Peel Kiwi, Peel Banana, Peel Onion
  - 5 videos for each test task for retrieval
  - Pair with 2495 randomly selected distracting videos

# Experiments - Datasets

- **COIN** (**CO**mprehensive **IN**structional video analysis)
  - 180 tasks, 11,827 videos
  - Unseen tasks, e.g., “Blow Sugar”, “Make Youtiao”, etc.
  - 5 test videos for each tasks, 900 in total

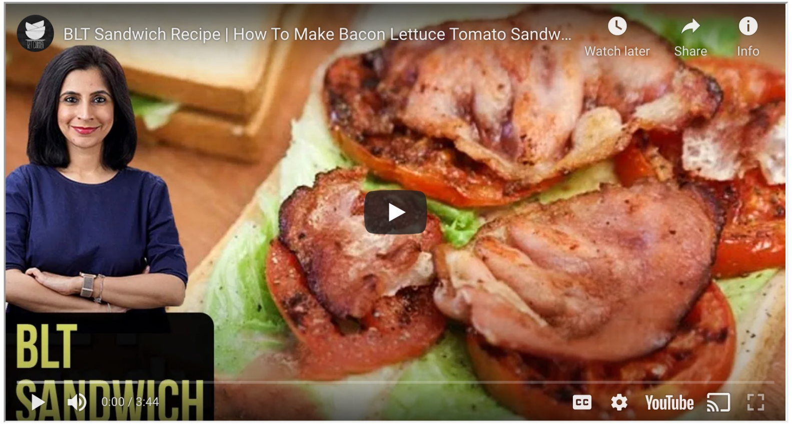



For each video, we annotate a series of steps with their temporal boundaries

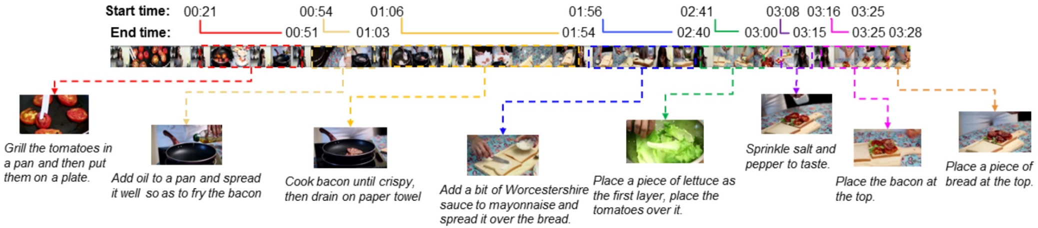
# Experiments - Datasets

- **Youcook2**
  - 89 recipes – tasks, 2,000 long videos
  - A retrieval pool of 436 videos (no overlap with Howto100M)




| Dataset | # of tasks | # of videos | Avg. video length (s) |
|---|---|---|---|
| Howto-GEN | 2,365 | 11,825 | 392.9 |
| COIN | 180 | 900 | 143.2 |
| Youcook2 | 89 | 436 | 310.9 |

Table 2. Statistics of the evaluation datasets (test set).



# Experiments - Baselines

- **Generation Models**
  - **T5** (Lyu et al. 2021)
  - **GPT-2-*large***
  - **GPT-3: Zero-shot generation**
    - **How to [Task Name]? Give me several steps.**
- **GOSC** (Lyu et al. 2021)
  - Goal-Oriented Script Construction
  - Step Inference model
  - Given the input task name
  - Gather the set of desired steps from wikiHow
- **Oracle**
  - Howto-GEN (from wikiHow)
  - COIN/Youcook2 (Human annotation)

Task: Make Tea

**GPT3**

- Steep tea leaves for 3-5 minutes.
- Pour tea into cups.
- Pour boiling water into the teapot.
- Put tea leaves into a teapot.
- Add sugar or honey to taste.

**GOSC**

- Find a tea you enjoy.
- Submerge tea leaves in boiling water.
- Select the tea varieties.
- Steep tea leaves in hot water
- Build a tea garden.

**Oracle**

- add some ingredients to the tea
- prepare and boil water
- pour the tea into the vessel
- prepare and add the tea
- heat the teapot and wash the cup
- add some water to the tea

# Results

| | Method | Howto-GEN | | | | | COIN | | | | | Youcook2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P@1↑ | R@5↑ | R@10↑ | Med r↓ | MRR↑ | P@1↑ | R@5↑ | R@10↑ | Med r↓ | MRR↑ | P@1↑ | R@5↑ | R@10↑ | Med r↓ | MRR↑ |
| | MIL-NCE [31] | 45.2 | 31.0 | 43.1 | 15.0 | .198 | 48.3 | 37.1 | 52.8 | 9.5 | .227 | 27.0 | 18.2 | 26.5 | 32.0 | .126 |
| Step Aggregation | T5 [30] | 44.0 | 29.9 | 41.0 | 19.0 | .190 | 46.1 | 35.3 | 50.7 | 10.0 | .219 | 21.3 | 16.0 | 24.7 | 61.5 | .108 |
| | GPT-2 [39] | 46.0 | 31.5 | 43.3 | 16.0 | .200 | 48.9 | 39.2 | 53.4 | 8.0 | .233 | 31.5 | 19.0 | 27.3 | 44.5 | .130 |
| | GPT-3 [2] | 49.3 | 33.3 | 45.7 | 13.0 | .211 | 53.3 | 42.1 | 59.0 | 8.0 | .252 | 37.1 | 22.4 | 34.6 | 27.0 | .160 |
| | GOSC [30] | 54.7 | 37.0 | 49.8 | 11.0 | .231 | 53.9 | 41.6 | 55.1 | 8.0 | .248 | 30.3 | 20.7 | 34.8 | 28.0 | .146 |
| | wikiHow | 51.9 | 35.4 | 47.8 | 11.0 | .222 | 53.9 | 40.8 | 56.1 | 7.0 | .246 | 31.5 | 21.0 | 34.2 | 24.5 | .149 |
| | IER (Ours) | 54.4 | 37.3 | 50.1 | 10.0 | .231 | **57.2** | 42.2 | 57.8 | **7.0** | .256 | **41.6** | **25.8** | **38.8** | **20.0** | **.175** |
| | IER$^3$ (Ours) | **55.0** | **37.4** | **50.6** | **10.0** | **.234** | 56.1 | **42.3** | **59.1** | 8.0 | **.258** | 40.4 | 25.1 | **38.8** | **20.0** | .172 |
| | Oracle | 56.5 | 38.0 | 50.8 | 10.0 | .237 | 60.0 | 43.4 | 59.3 | 7.0 | .262 | 52.8 | 33.5 | 47.1 | 14.0 | .215 |

Table 3. Retrieval performance on Howto-GEN, COIN and Youcook2. Baselines include retrievals based on global matching, aggregation of steps generated from state-of-the-art language models, goal-oriented script construction (GOSC), and wikiHow. The Oracle upper bound contains human-written step labels for each task. Observe that our **IER** systems outperform the baselines across all metrics.

# Results

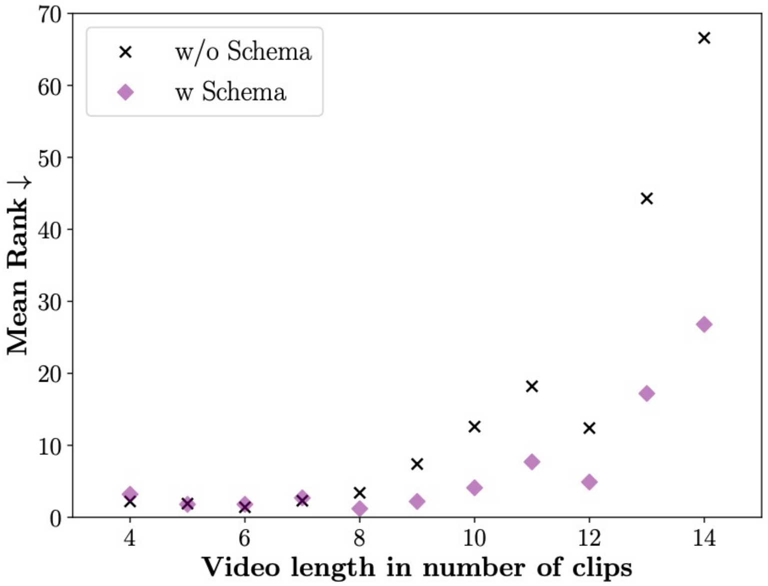



Figure 3. Retrieval performance by video length (in the number of clips). We group the test videos of Youcook2 by the number of clips per video and compute the mean rank for each group.

**IER helps more for longer videos**

# Editing Module Ablations

| | Method | P@1↑ | R@5↑ | R@10↑ | Med r↓ | MRR↑ |
|---|---|---|---|---|---|---|
| Howto-GEN | full | **54.4** | **37.3** | **50.1** | **10.0** | **.231** |
| | − mask | 53.7 | 36.3 | 49.3 | 11.0 | .229 |
| | − deletion | 53.6 | 36.9 | 49.8 | 11.0 | .230 |
| | − replacement | 51.5 | 34.9 | 47.3 | 12.0 | .220 |
| | − all | 45.5 | 31.0 | 43.1 | 15.0 | .199 |
| COIN | full | 57.2 | 42.2 | 57.8 | **7.0** | .256 |
| | − mask | 53.9 | **42.3** | 58.3 | **7.0** | .257 |
| | − deletion | **58.3** | 42.0 | 58.0 | **7.0** | **.258** |
| | − replacement | 53.8 | 41.0 | **59.2** | 7.5 | .251 |
| | − all | 54.4 | 39.6 | 53.7 | 8.0 | .246 |
| Youcook2 | full | **41.6** | 25.8 | 38.8 | **20.0** | **.175** |
| | − mask | 40.4 | 25.4 | 39.3 | **20.0** | .173 |
| | − deletion | **41.6** | **26.0** | 39.1 | 21.0 | **.175** |
| | − replacement | 40.4 | 25.8 | 38.5 | **20.0** | .173 |
| | − all | 40.4 | **26.0** | **39.9** | 21.0 | .174 |

Table 4. Ablation study on editing modules. "full" represents using all three modules and "− all" denotes removing all three modules. "− mask", "− deletion" and "− replacement" are short for removing "Token Replacement", "Step Deletion" and "Object Replacement" respectively. The numbers with underline are the ones lower than "full". The highest number of each metric is **bold**.

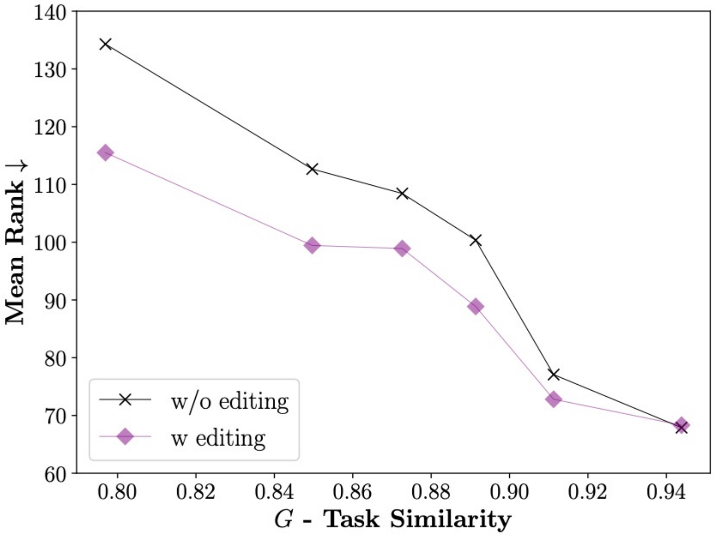



Figure 4. Retrieval performance by task similarity. We sort the test tasks of Howto-GEN based on their task similarity ($G$) and compute their mean rank for every batch of 400 tasks.

**Editing helps more when task similarity is low.**

# Schemata Transfer

- **Schemata can be reused by different video-text model**
- Use CLIP (Radford et al., 2021) as the video-text matching function
  - 400 million (image, text) pairs
  - Global Matching
  - Step Aggregation with schemata
- Schemata are transferable.

| Model | P@1↑ | R@5↑ | R@10↑ | Med r↓ | MRR↑ |
|---|---|---|---|---|---|
| MIL-NCE | 48.3 | 37.1 | 52.8 | 9.5 | .227 |
| +schema | 57.2 | 42.2 | 57.8 | 7.0 | .256 |
| CLIP [38] | 58.9 | 44.9 | 58.8 | 6.0 | .264 |
| +schema | 65.0 | 47.4 | 60.8 | 5.5 | .282 |

Table 5. Retrieval performance on COIN using MIL-NCE and CLIP as the matching functions. +schema represents using schema induced by IER (MIL-NCE as matching function) for retrieval.

# Conclusion & Future Work

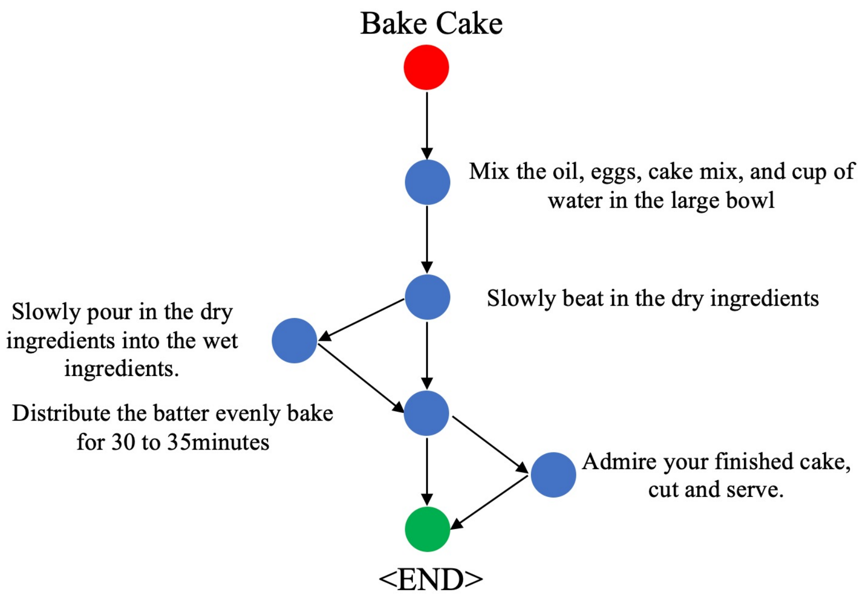
- We propose a schema induction and generalization system that improves instructional video retrieval performance.

- We demonstrate that the induced schemata benefit video retrieval on unseen tasks, and our IER system outperforms other methods.

- In the future, we plan to investigate the structure of our schemata.

# Conclusion & Future Work

- Temporal order in schema graph




- Other schemata applications
  - Video Anticipation
  - Task Identification
- Other aspects of schema
  - argument, duration, etc.
- Schema induction on other types of videos
  - News, human activities
  - Ego4D

# Thank you!