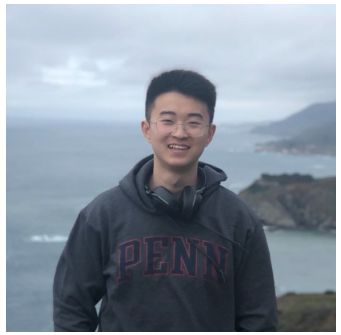
Induce, Edit, Retrieve: Language-grounded Multimodal Schemata for instructional Video Retrieval

Yue Yang



Joongwon Kim

Artemis Panagopoulou

Mark Yatskar

Chris Callison-Burch

Department of Computer and Information Science
University of Pennsylvania

Motivation

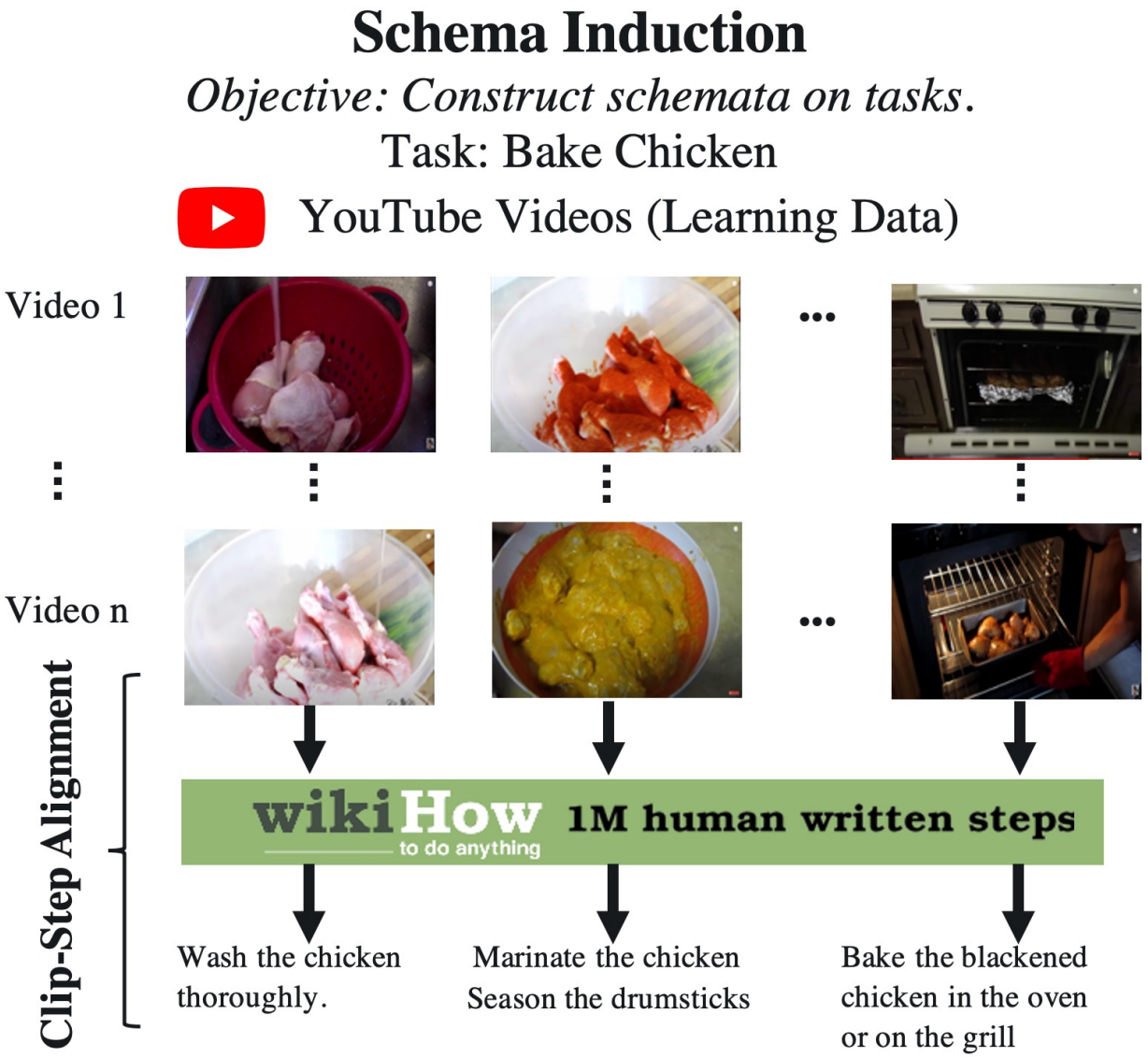
- Schema: a set of rules people use to perform everyday tasks.
- Schema can be generalized.
- When facing new tasks, people use prior knowledge. (Chen et al., 2004)

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



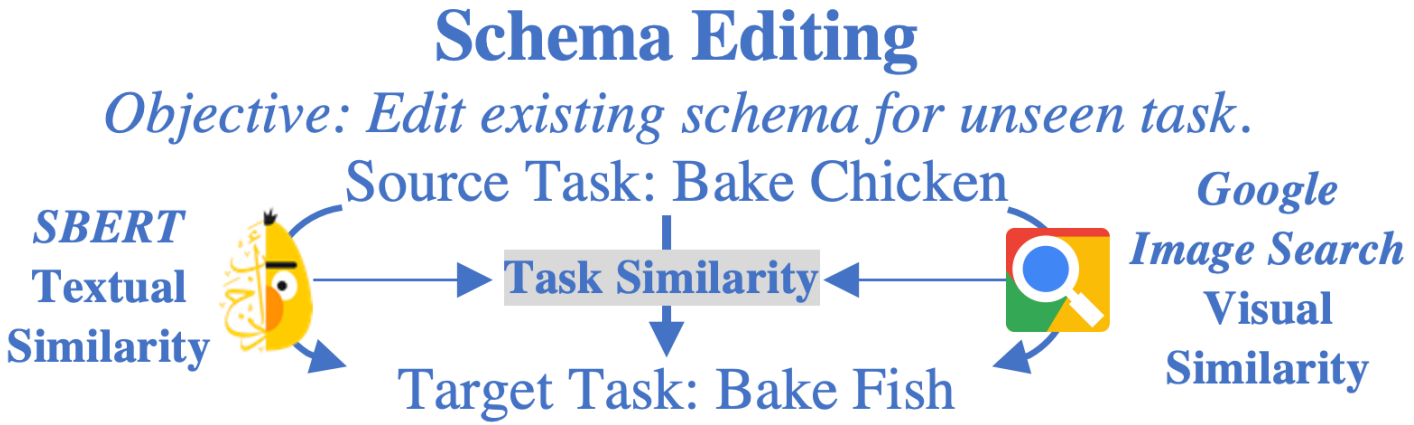
Schema of *Bake Chicken*

- 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
- … …

- Align wikiHow steps with videos.
- MIL-NCE as the video-text model.
- We induced 21,299 schemata using 1.2 M videos from Howto100m.

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

- 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 Put the ham in the oven. ↓ Put the lamb in the oven.
Clean a Guitar $\xrightarrow{0.84}$ Build a Violin Use a polish for particularly dirty guitars. ↓ Use a polish for particularly dirty violins.
Trap a Rat $\xrightarrow{0.84}$ Trap a Rabbit Bait and set snap rat traps. ↓ Bait and set snap rabbit traps.

Schema Editing

- 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 QA 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. ↓delete ~~Fill your pot with a balanced fertilizer.~~
Fix a Toilet $\xrightarrow{0.85}$ Remove a Toilet Test out the new flapper. ↓delete ~~Test out the new flapper.~~
Brush a Cat $\xrightarrow{0.87}$ Brush a Long Haired Dog Comb and groom your pet. ↓include Comb and groom your pet.

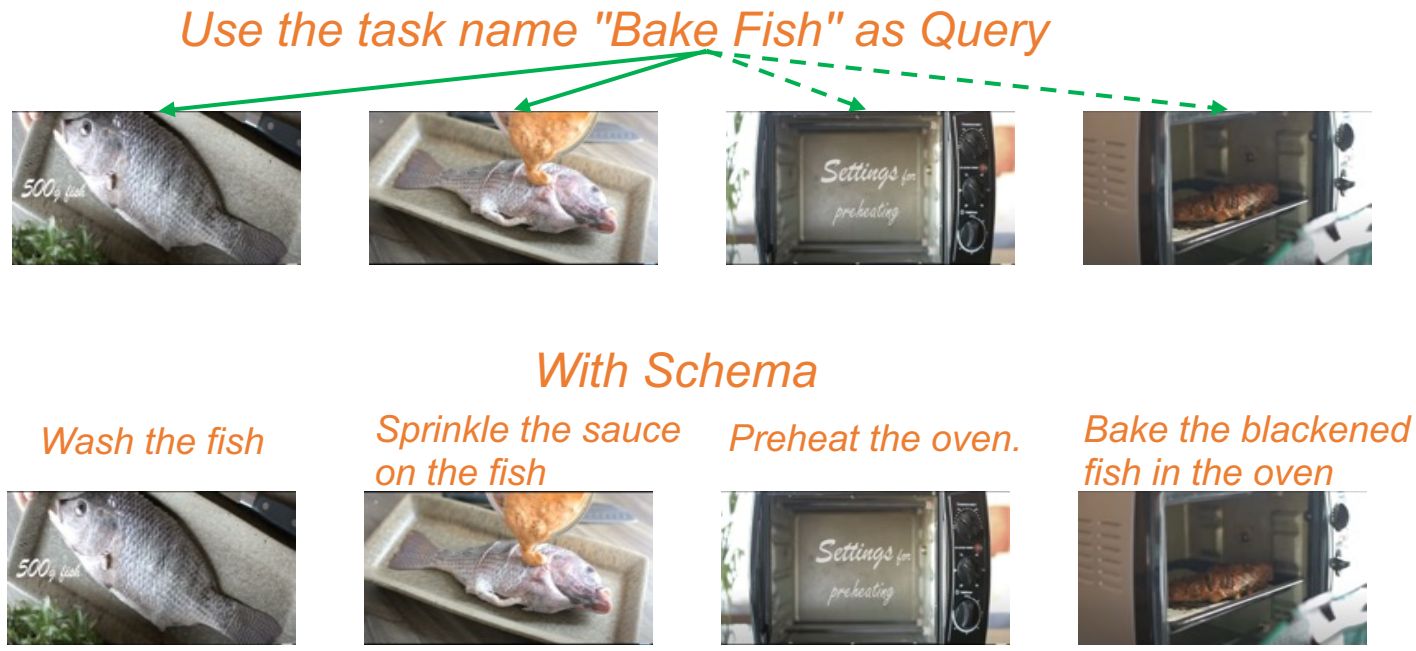
Schema Editing

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

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)



Experiment-Datasets

- **Howto-GEN** (a new split of Howto100M)
  - 3365 tasks, 2,184 unique main objects.
  - Random select 500 tasks for training, 500 for validation, 2365 for test.

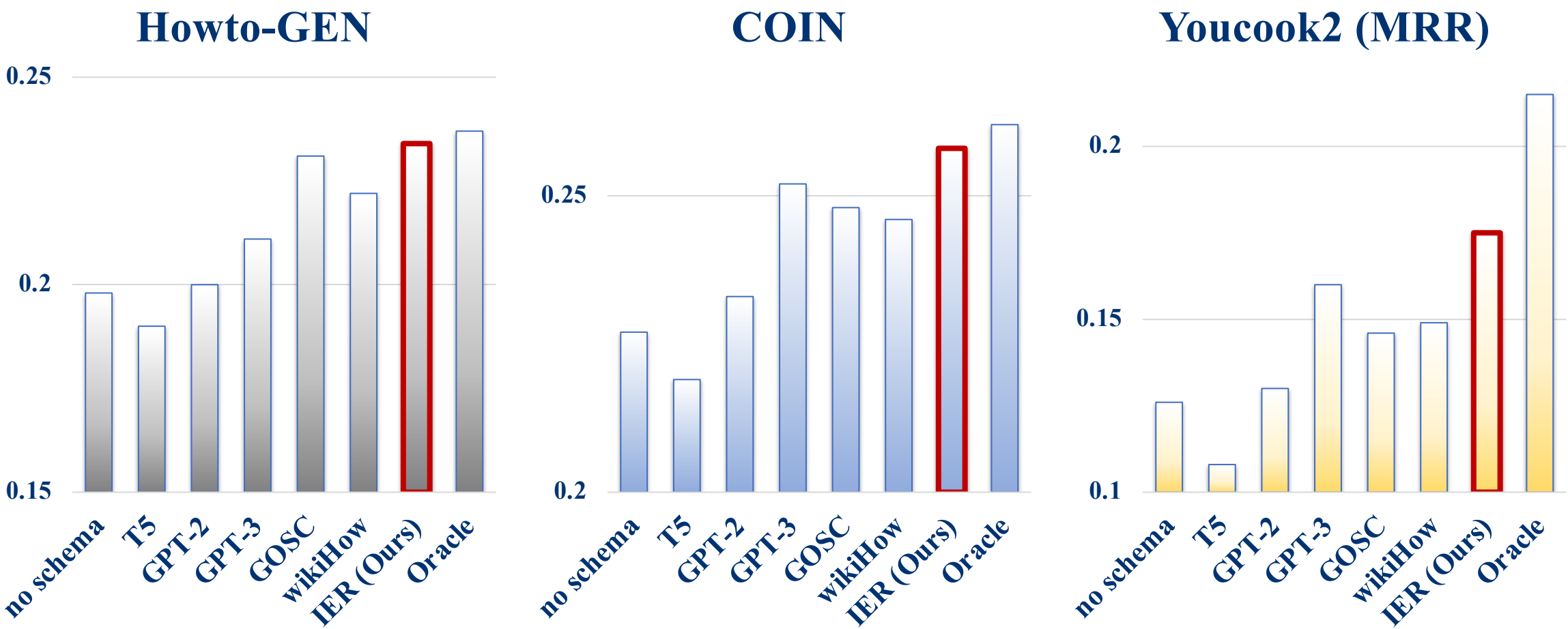
- **COIN** (**CO**mprehensive **IN**structional video analysis)
  - 180 tasks, 11,827 videos.
  - Unseen tasks, e.g., “Blow Sugar”, “Make Youtiao”, etc.

- **Youcook2**
  - 89 recipes – tasks, 2,000 long videos.

Experiment - Baselines

- **Generation Models**:
  - T5, GPT-2-large.
  - GPT-3: Zero-shot generation (How to [Task Name]? Give me several steps.)
- **G**oal-**O**riented **S**cript **C**onstruction
  - Given the input task name, retrieve the set of desired steps from wikiHow.
- Oracle Schemata (human written, upper bound)
  - Howto-GEN (from wikiHow)
  - COIN/Youcook2 (Human annotation)

Results:


Howto-GEN
0.25
0.2
0.15
no schema
T5
GPT-2
GPT-3
GOSC
wikiHow
IER (Ours)
Oracle
COIN
0.25
0.2
no schema
T5
GPT-2
GPT-3
GOSC
wikiHow
IER (Ours)
Oracle
Youcook2 (MRR)
0.2
0.15
0.1
no schema
T5
GPT-2
GPT-3
GOSC
wikiHow
IER (Ours)
Oracle

Results:

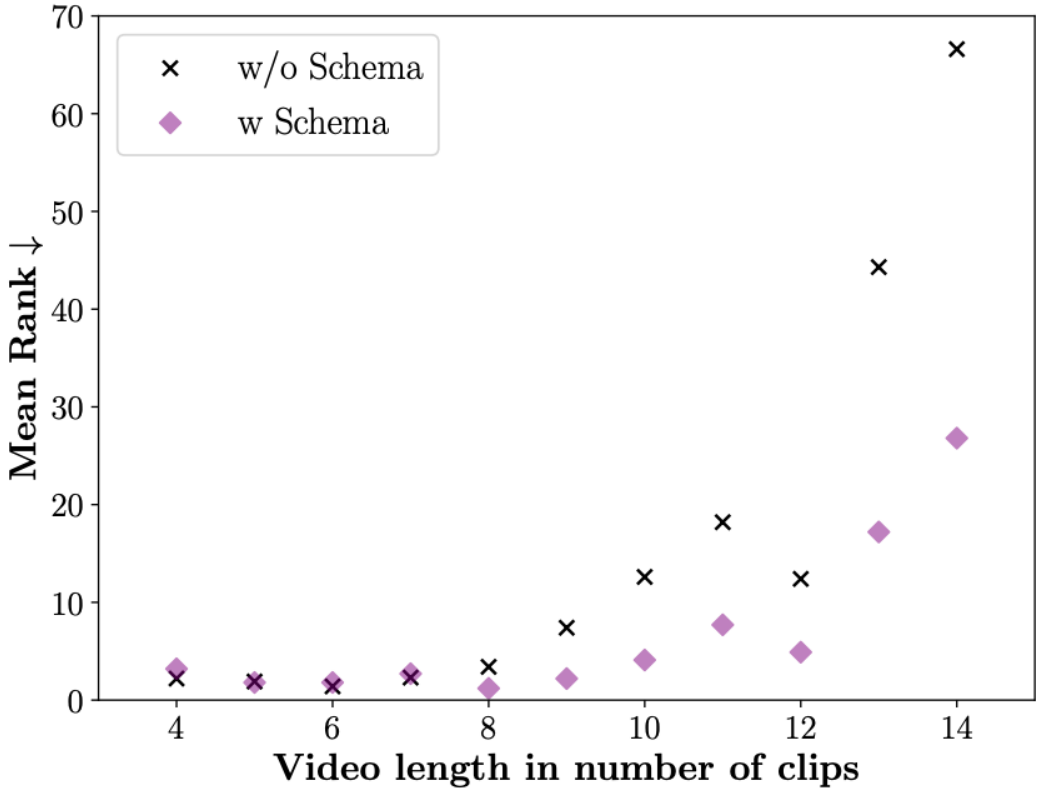


Figure 3. Retrieval performance by video length (in the number of clips). We group the test videos of Youcook2 by the number of clips per video and compute the mean rank for each group.

IER helps more for longer videos

Results:

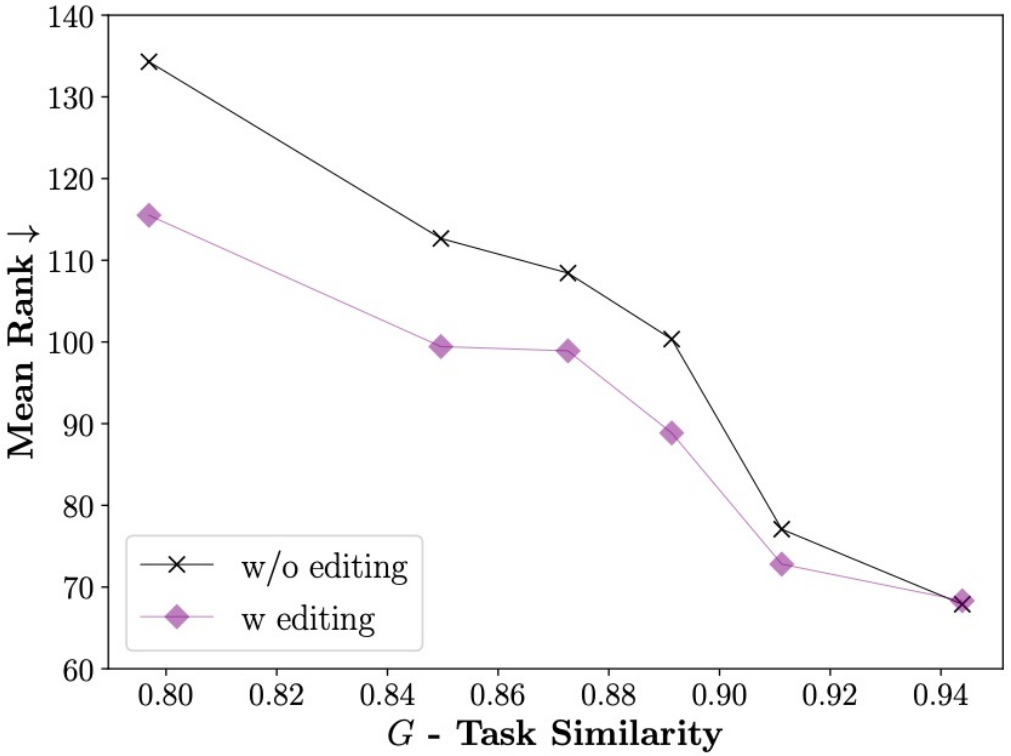


Figure 4. Retrieval performance by task similarity. We sort the test tasks of Howto-GEN based on their task similarity (G) and compute their mean rank for every batch of 400 tasks.

Editing helps more when task similarity is low.

Results:

	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

All three editing modules are beneficial.

Results:

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.

Schemata can be reused by different video-text model.

Conclusion

- 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.

Thank you!