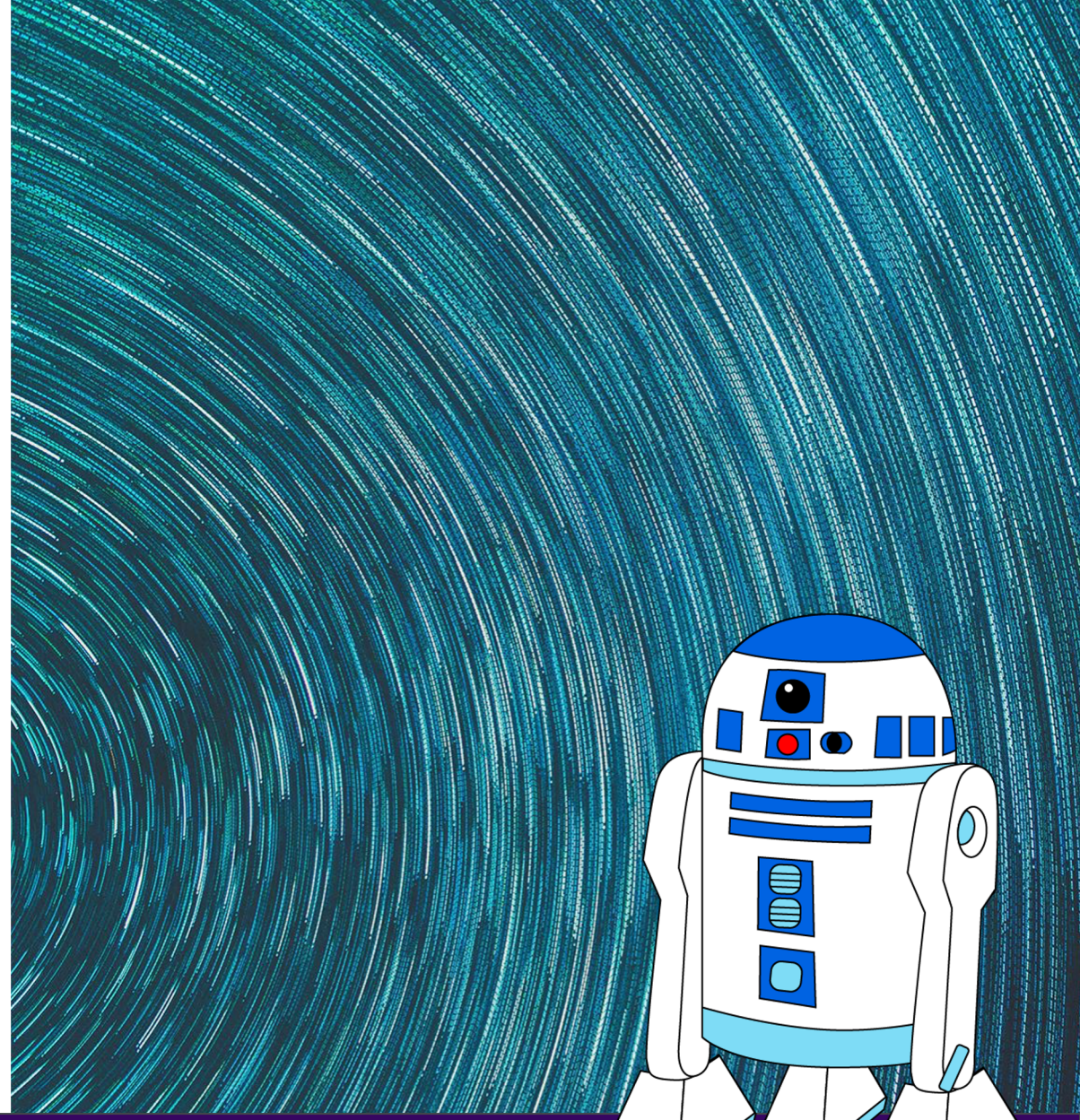
wikiHow + Video

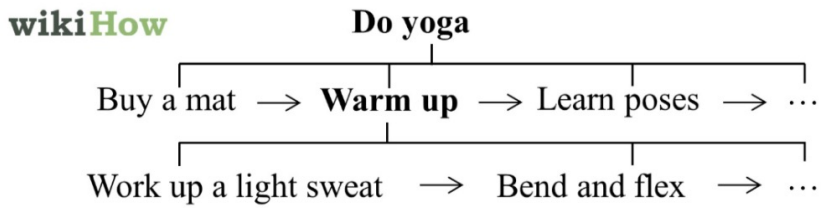
Multi-modal Schema Induction

Presenter: Yue Yang



Penn Engineering
University of Pennsylvania

Motivation



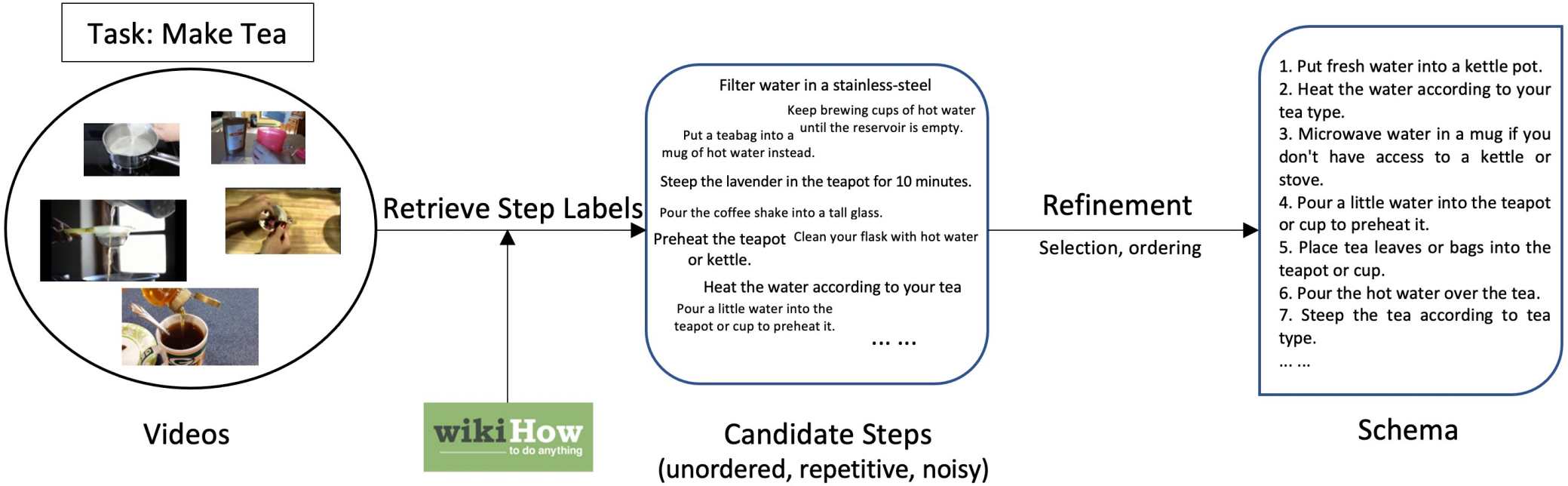
- ***Schema - high-level representations of complex events***
  - Narrative Cloze Test (Chambers and Jurafsky, 2008)
  - Language modeling variations (Rudinger et al., 2015)
  - Event temporal ordering, goal->step inference, step->goal inference, clustering, next-event prediction (Ragneri et al., 2010; Zhang et al., 2020)

- ***Problems:***
  - Manually-defined schema is **labor-intensive**
  - Low coverage and fail to generalize to **new domains**
  - Previous methods induce schema from document (**textual data**) only

- ***Our Solution:***
  - **Video** contains temporal information (low-level actions) to accomplish complex task
  - Induce schema from a large number of videos describing the same task

Main Idea

- Input: a bag of videos (describing the same task)
- Output: schema (ordered set of steps representing the high-level goal)
- Two stages:
  1. Retrieve step labels for clips
  2. Refinement (selection, ordering)



Problem Formulation

1. Given **a high-level goal** ***G*** with corresponding **videos** $V = \{v_1. v_2, v_3, \dots, v_n\}$.
2. **Segment** these videos into short **clips** $C = \{c_1, c_2, c_3, \dots, c_m\}$.
3. **Compare** these clips a huge set of **existing text steps** $S = \{s_1, s_2, s_3, \dots, s_M\}$.
4. Collect a super set of potential steps $S' = \{s'_1, s'_2, s'_3 \dots, s'_m\}$.
5. **Filter** and **order** these steps to formulate the final schema.

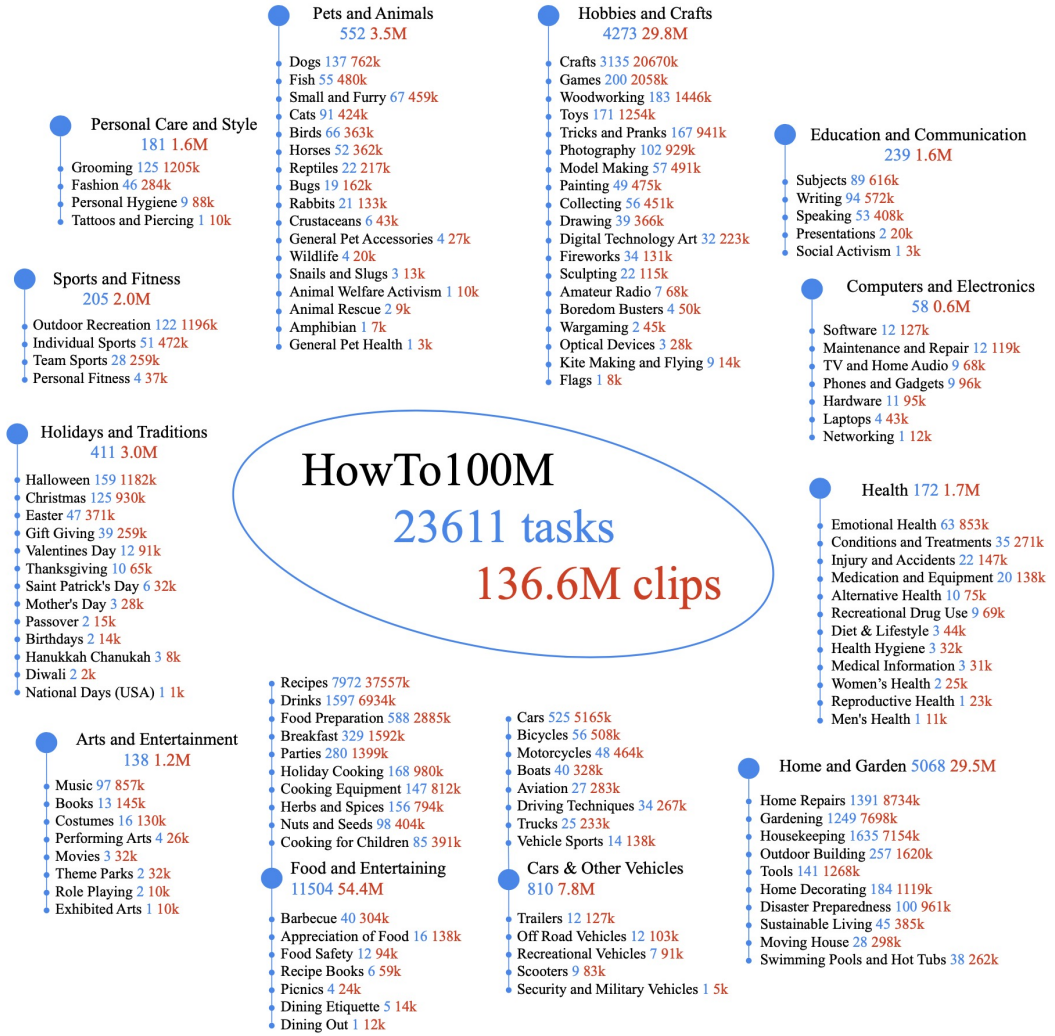
Retrieve

→ Refinement

Dataset

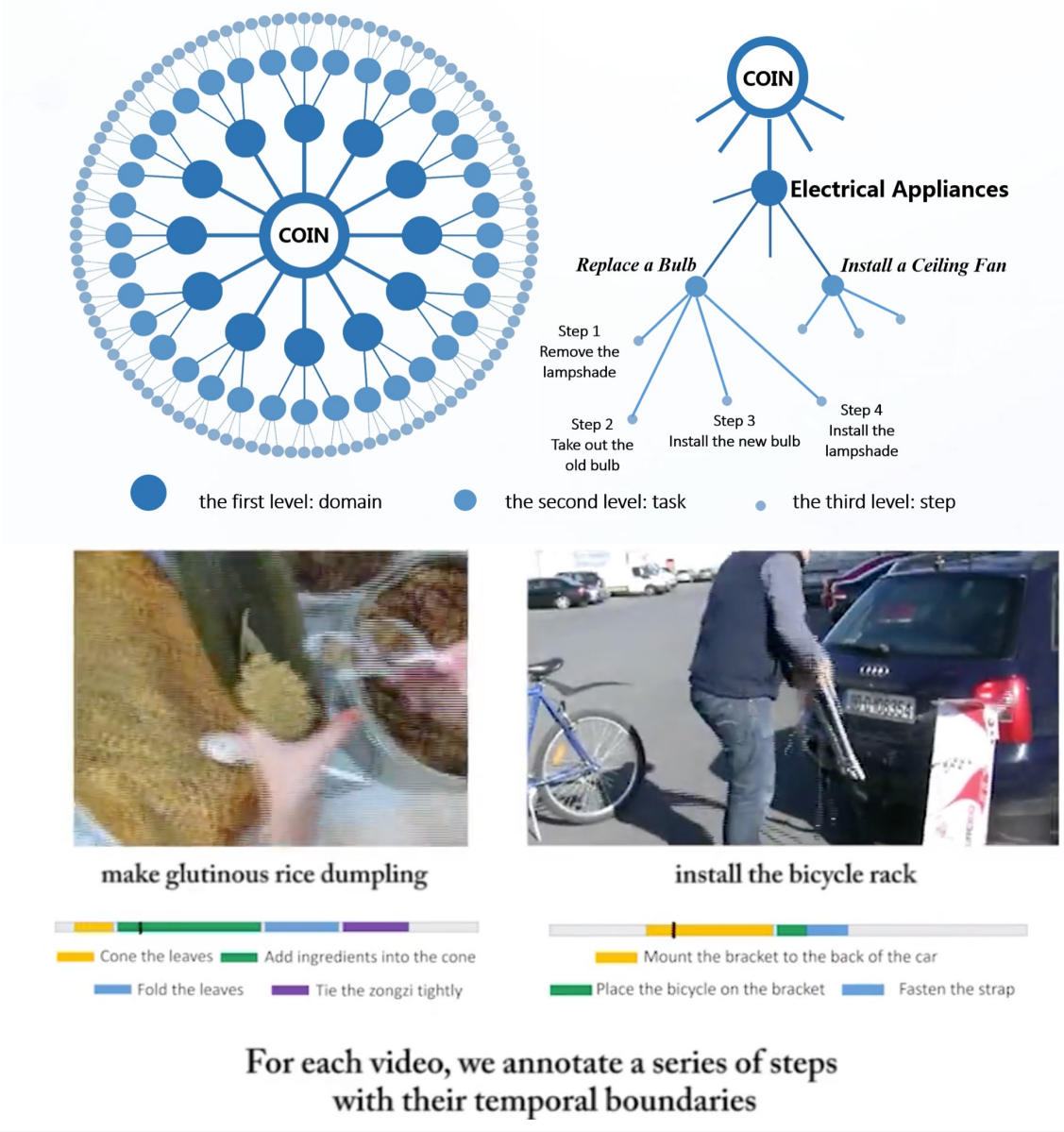
- **wikiHow**
  - **120K** professionally edited how-to articles
  - spanning a wide range of domains
  - Goal-step structure
  - Over **110M** steps

- **Howto100M** (Miech et al., 2019)
  - **136 M** video clips from **1.22M** narrated instructional videos from **23K** tasks
  - Tasks are directly from wikiHow
  - Focus on visual tasks only
  - Retrieve videos from YouTube
  - Construct (clip, caption) pairs use **A**utomatic **S**peech **R**ecognition (ASR)
  - In Domain



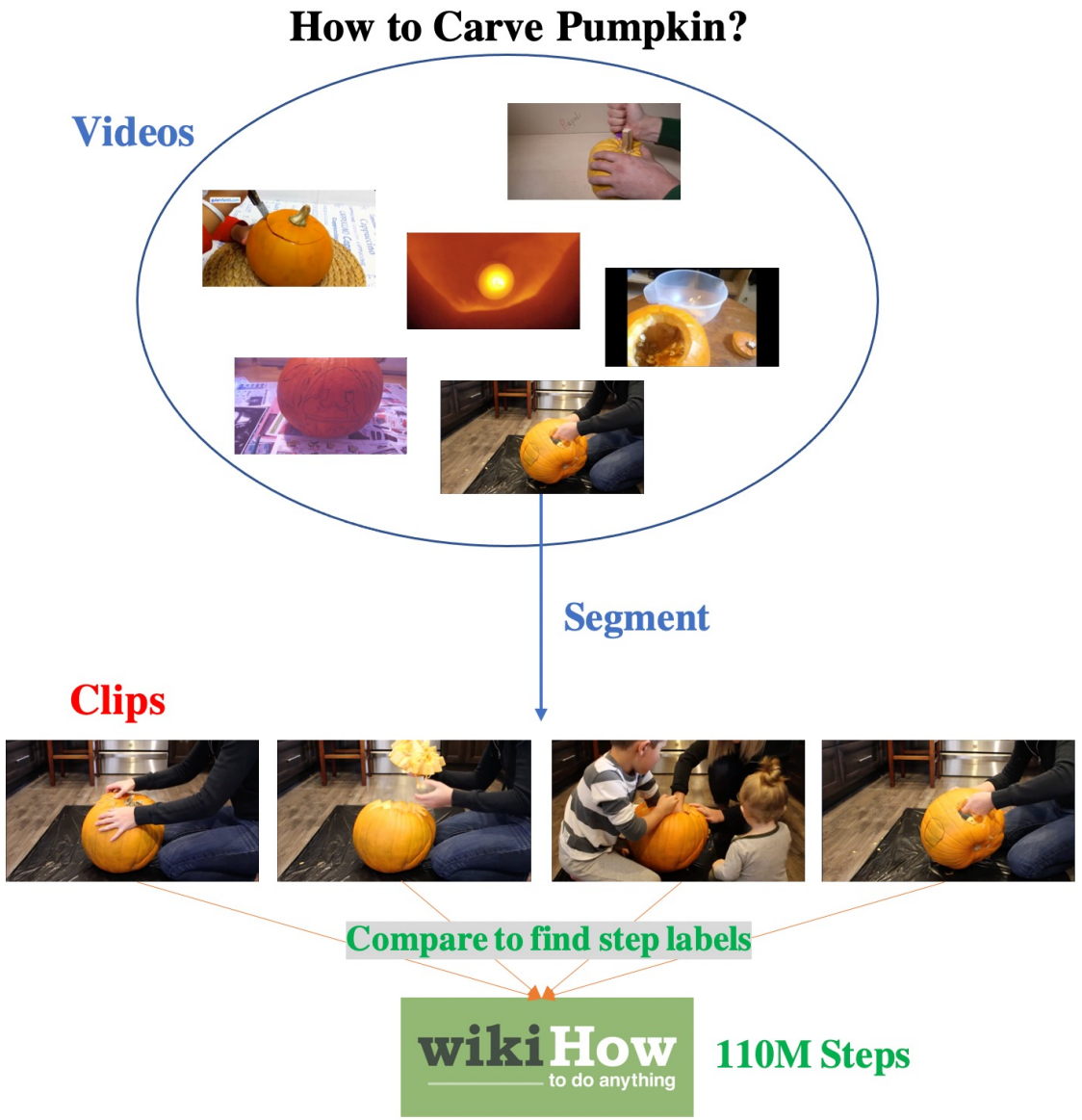
Dataset

- **COIN** (Tang et al., 2019)
  - hierarchical structure (Domain-Task-Step)
  - **11,827** videos related to **180** different tasks
  - Collected from YouTube
  - **Human Annotation (Step labels):**
    - Given pre-defined step labels
    - Select (star, end) for the step
    - Each video is labelled with **3.91** step segments
    - each segment lasts **14.91** seconds on average
  - Out-of-domain



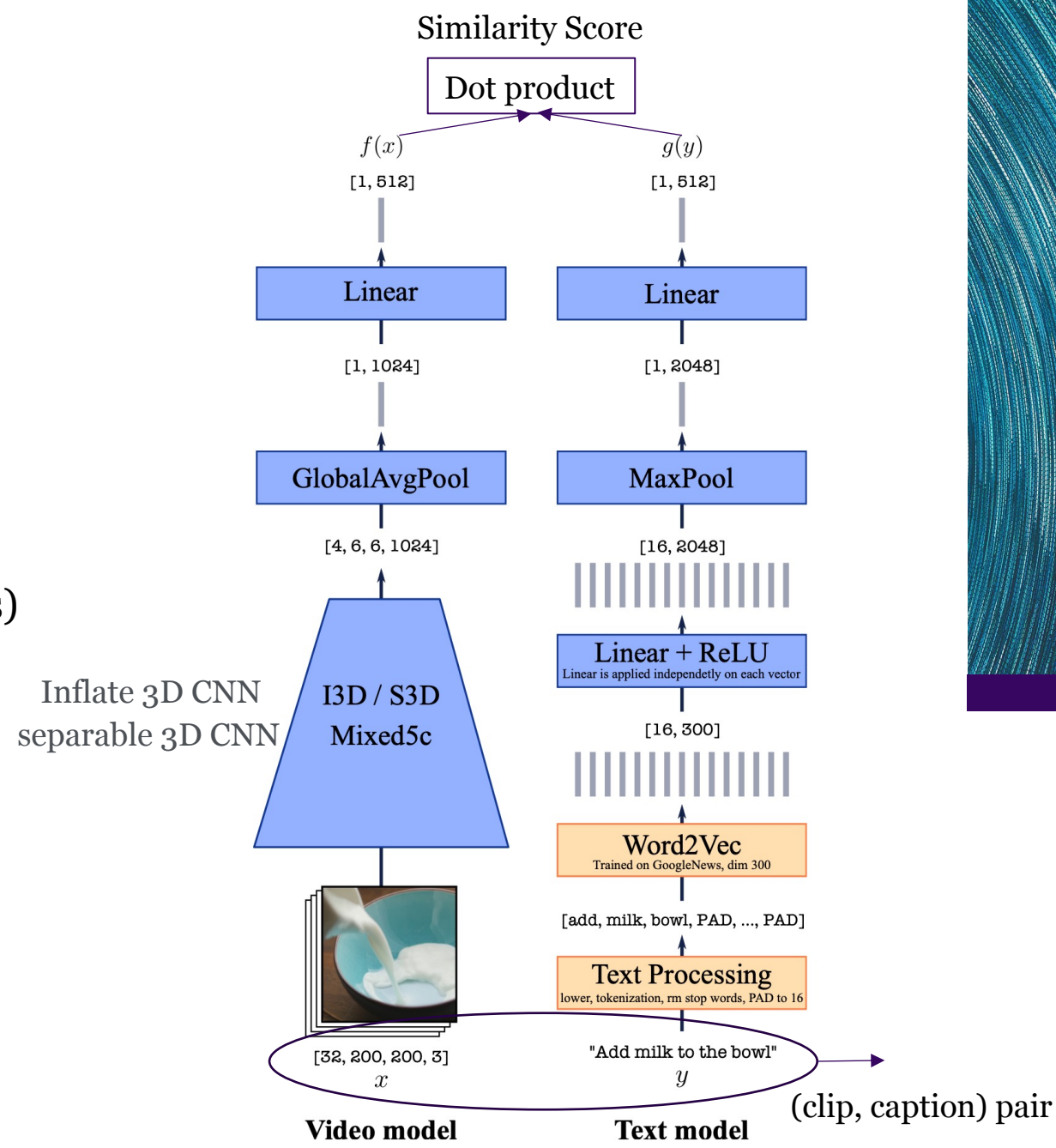
Retrieve Step Labels

- **Input:** collection of videos (carve pumpkin)
- Segment videos into clips
  - Howto100m: ASR (110 clips / video)
  - COIN: annotated clips (4 clips / video)
- For each clip, compare with **all** existing steps in wikiHow
- **Multimodal comparison:** pre-trained model



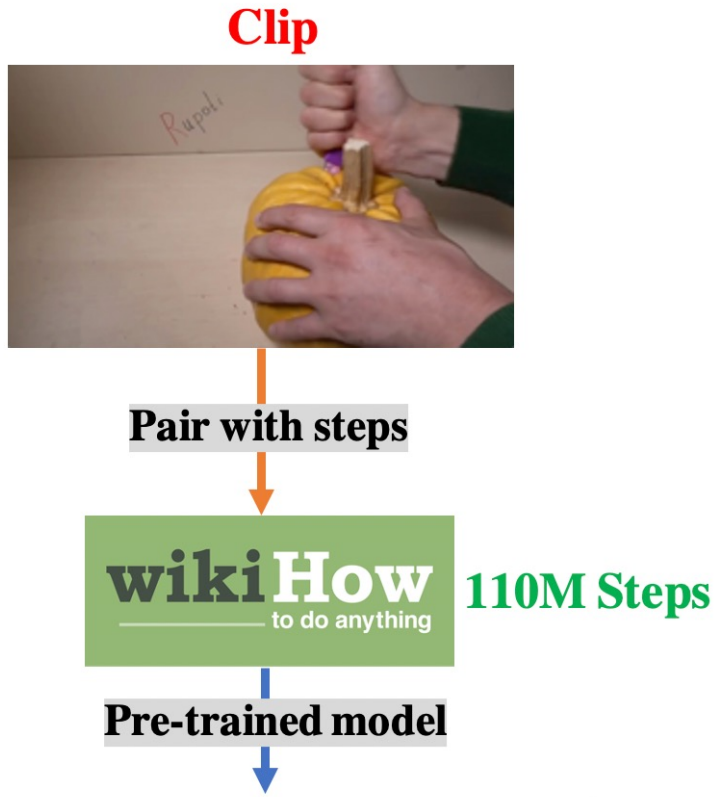
Retrieve Step Labels

- **Pretrained model**: MIL-NCE (Miech et al., 2020)
  - Multiple Instance Learning (MIL)
  - Noise Contrastive Estimation (NCE)
- Training on **Howto100m**
- **Supervision**: (clip, caption) pair from ASR
- **Video model**
  - Input a video (frames, height, width, channels)
  - Output vector (1, 512) in joint space
- **Language model**
  - Word embedding with linear layer
  - Output vector (1, 512) in joint space
- **Similarity score**
  - **Dot product** the feature vectors of two modalities



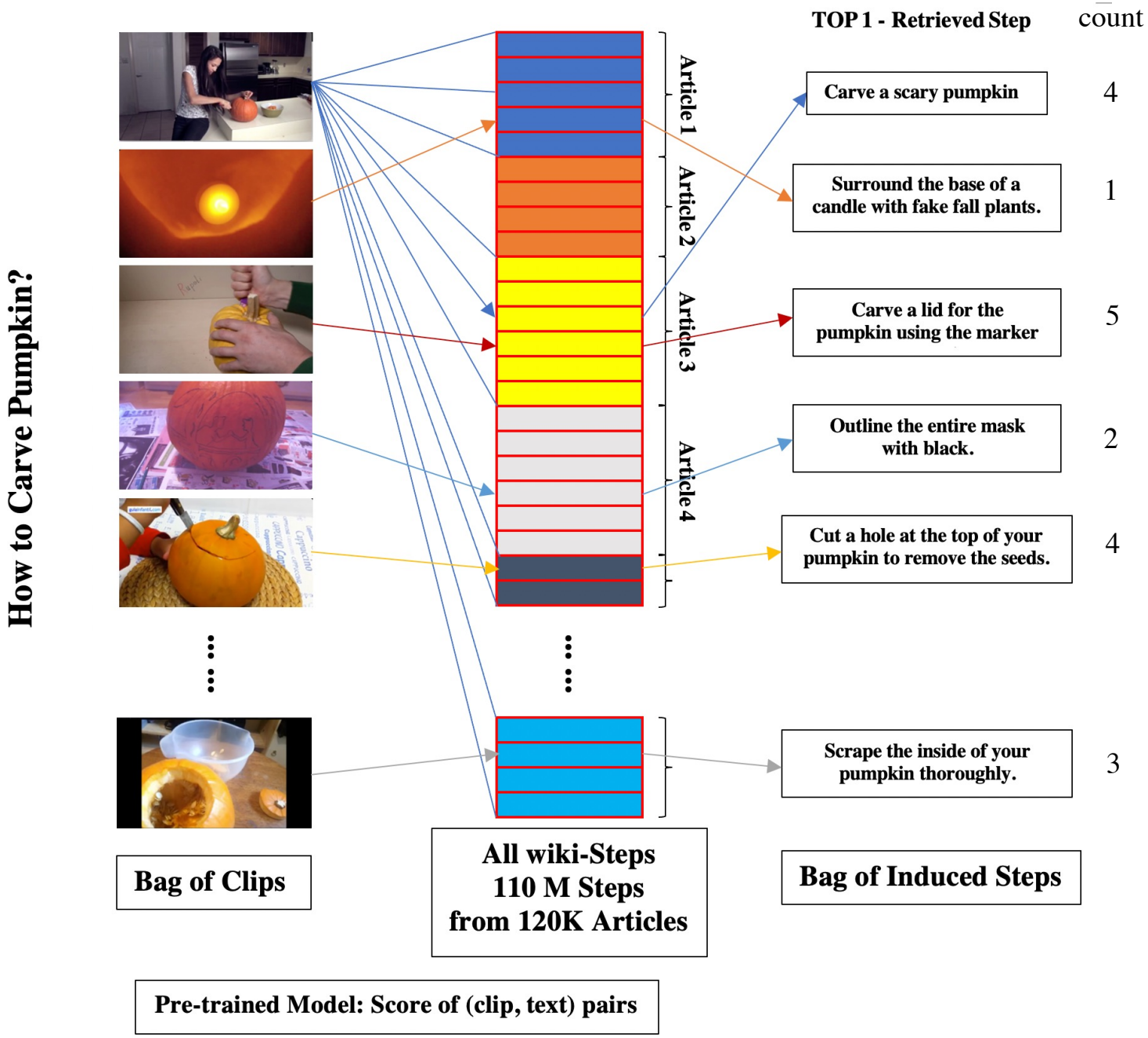
Retrieve Step Labels

- For each clip, pair with **all wiki-steps**
- Compute the similarity scores for **every pair** via pretrained model
- **Rank** these steps based on the scores
- Conduct this operation on all clips



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

Retrieve Step Labels



Retrieve Step Label – COIN

Task: Make Tea
Retrieved step label: Strain the tea through a filter and pour it into cups.
Ground truth label: add some ingredients to the tea

Task: Carve Pumpkin
Retrieved step label: Scoop the seeds out of your pumpkin with a large serving spoon.
Ground truth label: clean up the interior of the pumpkin

Task: Use Jack
Retrieved step label: Jack the car up so that you can fit, comfortably, underneath the car.
Ground truth label: raise the jack up

Most Frequent Steps – Howto100m

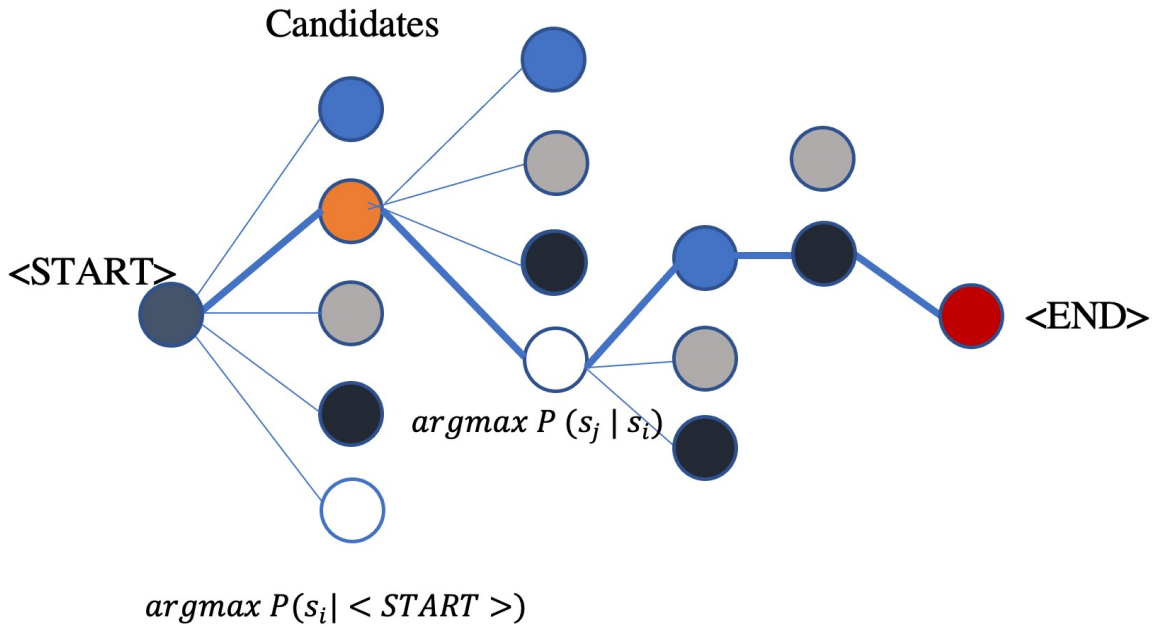
How to Stain Cabinets	How to Use a Drill Safely	How to Replace Shocks
(9) **Glaze** the doors using the same process you did with the cabinets. (4) Choose a whitewash **wood stain**. (4) **Paint** dated **cabinets** and dark walls. (4) Finish the **cabinets** with a **top coat**. (4) Apply **glaze** to a section of one **cabinet door** or drawer. (3) Opt for semi-custom cabinets for a midrange budget option with more features. (3) **Prime the cabinets** with white primer paint. (3) Put a lazy susan in your cabinets. (2) Choose an appropriate **urethane** finish for the door. (2) Apply the dye to the poplar with a **rag**.	(19) Set the **plunge depth** for the drill. (8) Put on **safety glasses** before you start drilling. (6) **Secure the cord grip** by installing the grub screw with an **Allen wrench**. (5) Wear **safety goggles** and **a dust mask** while drilling. (5) **Locate** the **chuck** at the end of the drill. (4) Drill your team with simulated data breaches. (4) Drill through the tile slowly. (4) Set up your guide rail for cutting with a plunge saw. (4) Complete **routing** and other machining before ebonizing. (4) Wear the proper **safety gear** when **sawing and drilling** into wood.	(13) Visually **inspect** your **strut mounts** or **shock towers**. (10) Call the bank's toll-free customer service number. (9) Sign up for an email service. (9) Drop it off at an **auto repair** or **auto parts shop**. (9) Replace each hubcap. (8) Inspect your **wheel** wells and **bumpers**. (8) Examine the lug nuts. (7) Take your vehicle to a **reputable repair shop** for **diagnosis and repairs**. (7) Keep your tires aired up. (6) Loosen the bleeder.

Examples of Howto100m
(Ordered by counting the number of times being selected as the **top-1** clip label)

Ordering Steps using Language Model

- **Step ordering task** (Zhang et al., 2020b,a)
  - Given two steps (step1, step2), predict whether step1 happens before step2
- **Use the pair-wise conditional probability (local maximum)**
  - Fine-tune a BERT on **N**ext **S**entence **P**rediction (**NSP**) task on original wiki-step order
  - Obtain the probability of p (step 2 | step 1)
  - **Greedily** iterate through the induced steps to generate the chain of steps using the pairwise potentials

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



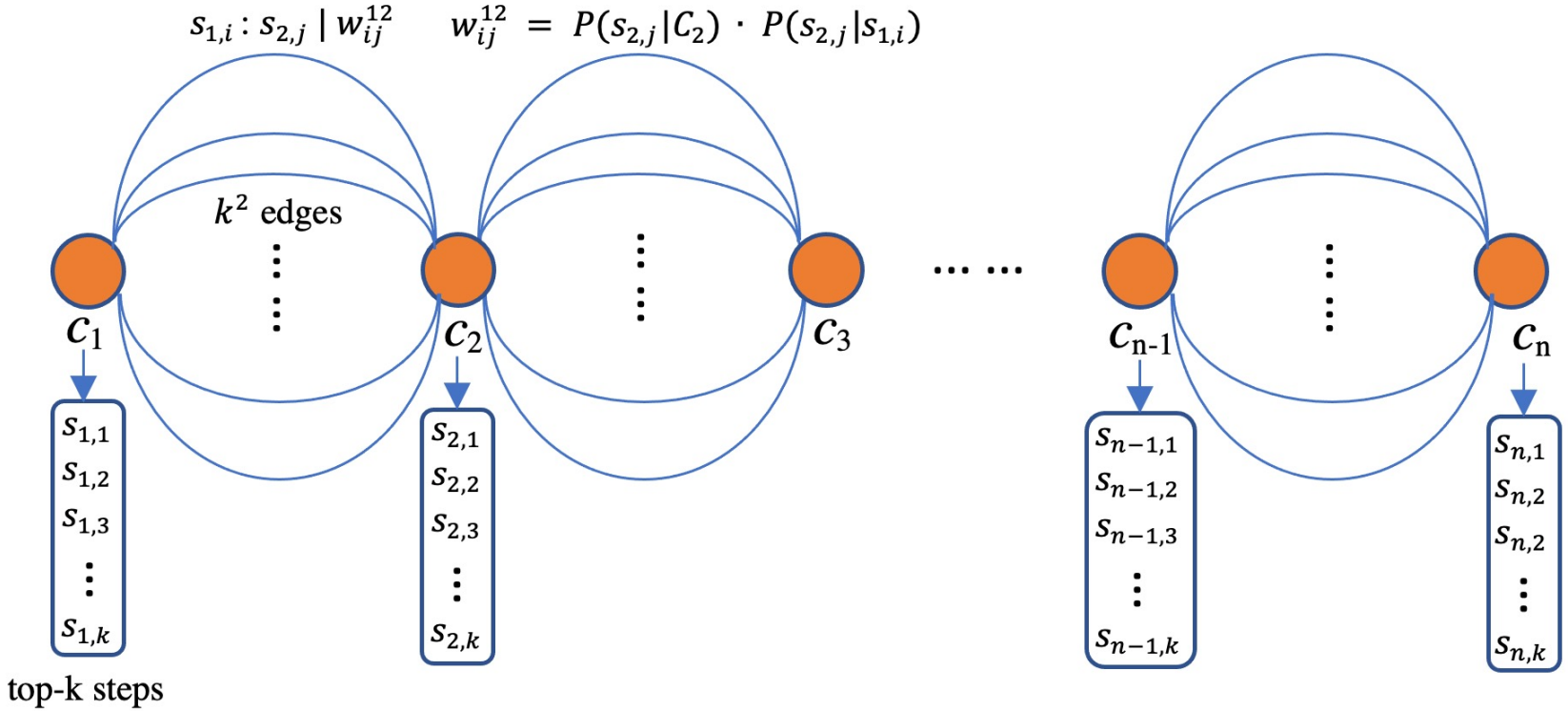
Problem: Ignore the **temporal information** in the video

Ordering Steps using Clip Order

- **Assume the order of steps follows the order of clips**
- **Constructing the Graph**:
  - Each step as a node
  - Go back to each video, draw a **directed edge** if we find one step happens after another step, e.g, (clip1: step A), (clip 2: step B) == step A --> step B
  - Assign the **weight** for the edges by **counting** the times that direction (order) happens
  - Result in a **D**irected **A**cyclic **G**raph (**DAG**) (remove cycles if necessary)
  - Find a **path** in the graph as the schema
- **Problem**
  - Ignore the **semantic relationship** between steps

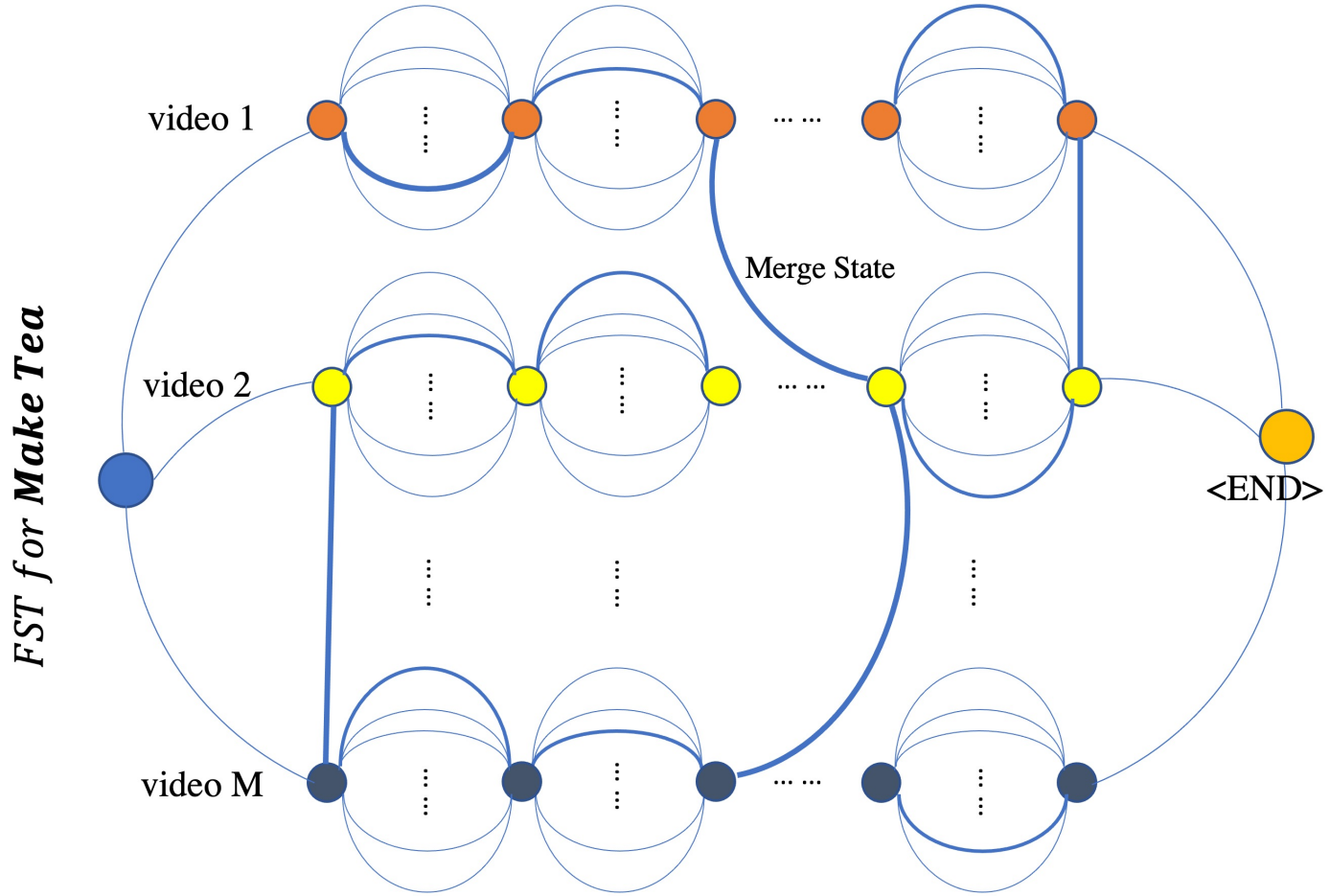
Ordering steps with LMs and clip order in videos

- **Finite-state transducer (FST)**
  - Each clip as a state, select top-k steps for each clip based on the step label retrieval
  - There will be k^2 edges between two states
  - Compute the weight of each edge by multiplying two probabilities (step retrieval, NSP)
  - Find the longest path (highest sum of weights)



Combining all videos into FST

- **Merge the states with high visual similarity**

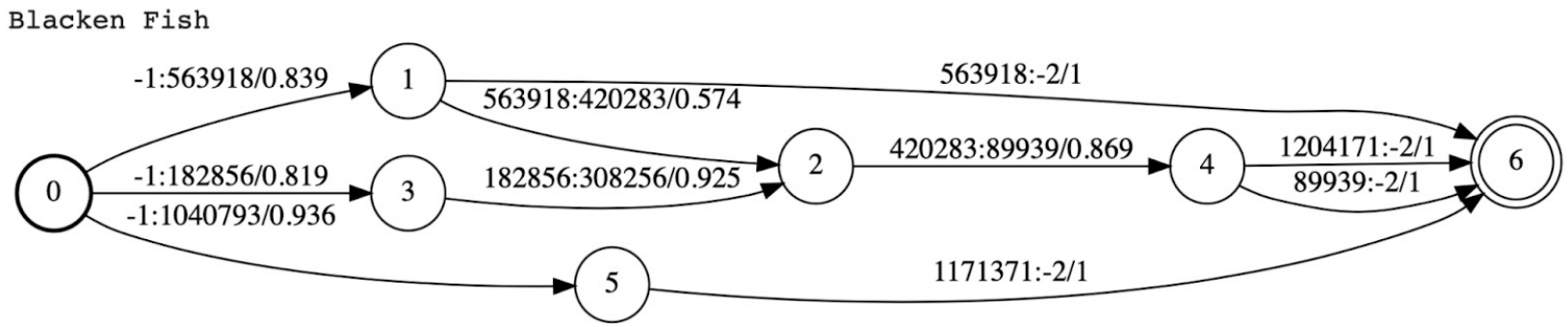


Video-FST

Stain Cabinets
• Glaze the doors using the same process you did with the c abinets. • Glaze your cabinets. • Finish the cabinets with a top coat.
• Choose a gel stain, glaze, or water-based stain if you want a darker finish. • Glaze the doors using the same process you did with the cabinets. • Pour liquid deglosser onto a cloth and rub it onto the doors and drawers. • Wipe in the direction of the grain. • Choose an appropriate urethane finish for the door.
• Choose an appropriate urethane finish for the door. • Consider applying some antiquing glaze for an aged look. • Apply the stripper to the cabinet doors and scrape off the gelled finish, as you did with the outside of the cabinets.

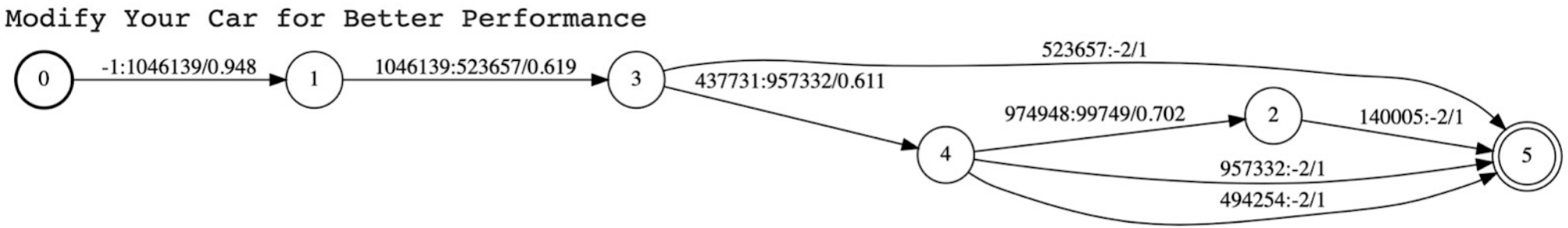
Use a Drill Safely
• Put on safety glasses before you start drilling. • Place your bit on the point where you'd like to drill and s queeze the trigger. • Pre-drill and paint your timber. • Hold the drill so the bit is perpendicular to the wood.
• Wear safety goggles and a dust mask while drilling. • Hold the drill so the bit is perpendicular to the wood. • Locate the chuck at the end of the drill. • Begin drilling at a steady speed. • Check the RPM on accessories against the RPM on the grinder before buying them.
• Put on safety glasses before you start drilling. • Secure the cord grip by installing the grub screw with an Allen wrench. • Begin drilling at a steady speed.

Task FST



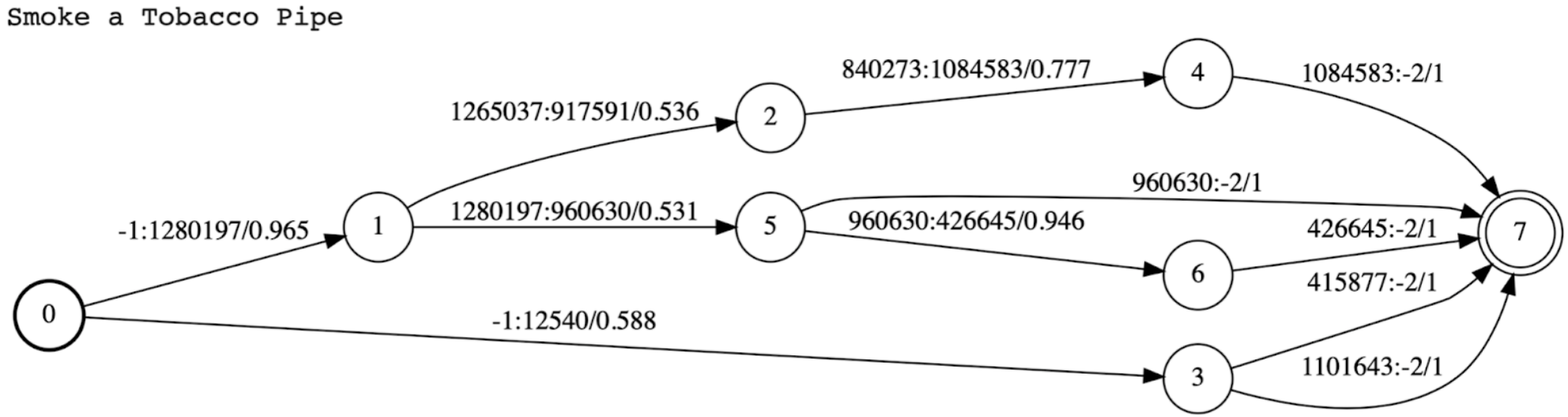
563918 : Sprinkle both sides of the meat with blackening spices.
182856 : Pat the fish dry and season it.
1040793 : Grill yellowfin tuna steaks.
420283 : Flip the fish and add a bit more butter.
89939 : Garnish and serve the lemon butter tilapia.
308256 : Turn and season the other side of the veal chops.
1204171 : Plate the fish and lemon, and wipe the skillet.
1171371 : Use tongs to take the oysters off of the grill.

Task FST



1046139 : Install a turbocharger kit.
523657 : Evolve Transmissions.
140005 : Mirror the movements of the instructor in the dance video.
437731 : Race go-karts.
957332 : Seek the advice of a mechanic.
974948 : Consult a mechanic if you have any questions.
99749 : Execute the hands-on training for the forklift operators.

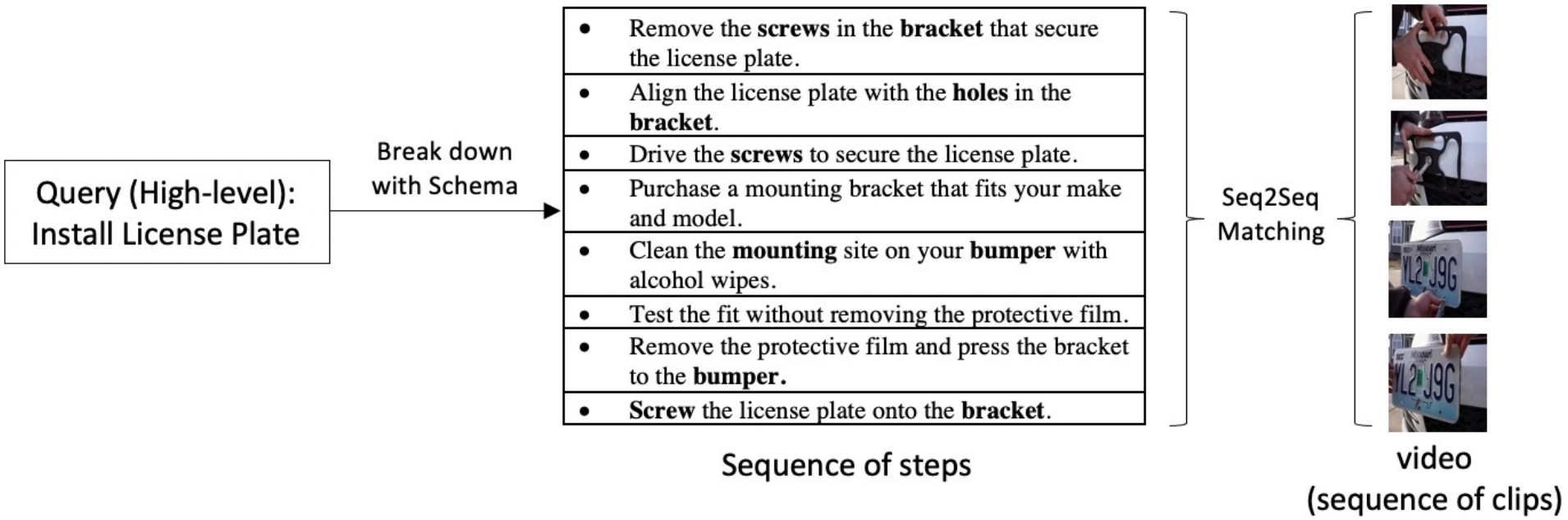
Task FST



1280197 : Clean the pipe after each smoke.
12540 : Smoke the lightest cigarettes you can.
1265037 : Try the baseball grip.
917591 : Throw out the tobacco.
960630 : Replace your flint on a Zippo lighter.
840273 : Give up tobacco.
1084583 : Try flavored cigarettes.
415877 : Make sure you're legally able to buy cigarettes.
1101643 : Ensure that the seller accepts PayPal.
426645 : Smoke your hookah.

Downstream Task – Schema-Guided Video Retrieval

- **Objective:** retrieve related videos given a **high-level** goal query
- **Problems: modality imbalance**
  - Lack of intermediate information in the textual query
  - Videos consist of sequence of detailed steps
- **Solution:** Use schema to break down high-level goal into low level steps (Seq2Seq Ma



Downstream Task – Schema-Guided Video Retrieval

- Dataset: COIN (videos are grouped by 180 high-level Tasks)
- Baselines:
  - 1. Match the video titles with the high-level prompt
  - 2. Match the goal only with the videos

$$S(G, V) = \max_{i=1:n} S(G, c_i)$$

- Schema guided model (unordered schema):

$$S^*(G, V) = (1 - \lambda_1) \cdot \underbrace{S(G, V)}_{\text{Goal score}} + \lambda_1 \cdot \frac{1}{n} \sum_{i=1}^{n} \underbrace{\max_{j=1:t} S(s_j, c_i)}_{\text{Step score}}$$

  - 1. Get the schema from the most similar article in wikiHow (match goal with article titles)
  - 2. Use the induced schema
- Upper bound: use the step labels from COIN annotated by human as the schema

Downstream Task – Schema-Guided Video Retrieval

		Goal-Video Retrieval (900 Videos)				
	Model	R@1 ↑	R@5 ↑	R@10 ↑	R@25 ↑	Mean r ↓
	Title (Bert)	**14.11**	44.33	56.33	70.00	52.91
Content-based	Goal only	12.11	45.33	60.78	74.11	38.34
	Goal + Step (original)	12.11	46.67	63.11	78.56	32.10
	Goal + Step (hierarchical)	12.11	45.55	63.22	77.67	32.44
	Goal + Step (induced)	12.89	48.22	66.56	**83.44**	**18.06**
	Goal + Step (GT)	13.78	**50.78**	**67.44**	82.33	24.89

Table 8: Retrieval Performance on COIN Testing Set (GT for ground truth)

- Content-based model is more robust (use title only depends on the quality of the titles)
- Use schema to break up goal could improve performance
- Induced schemas outperform the manually-defined schemas (even the GT)

Thank you!