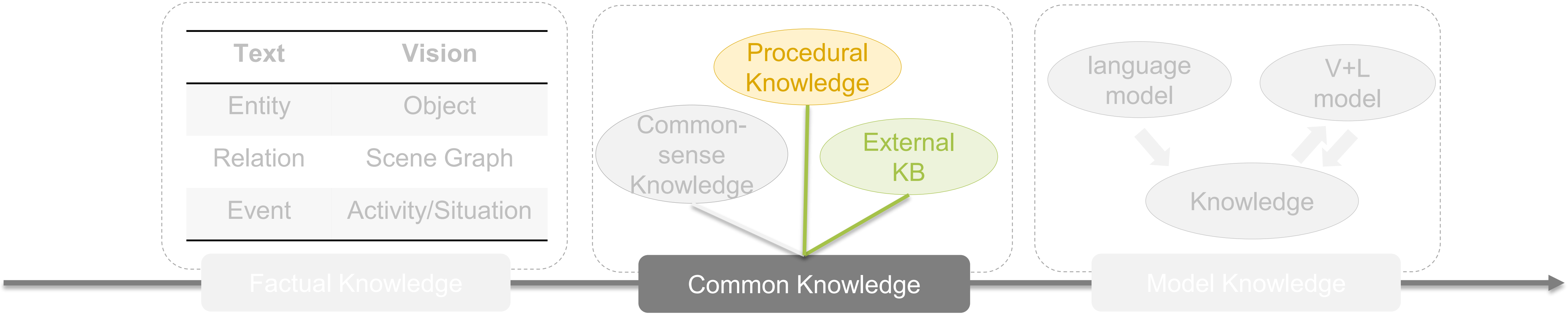
Procedural Knowledge

Knowledge-Driven Vision-Language Pretraining (Part IV)

Xudong Lin
Columbia University
xudong.lin@columbia.edu

Content

Learning patterns of procedure with human-curated patterns and data.

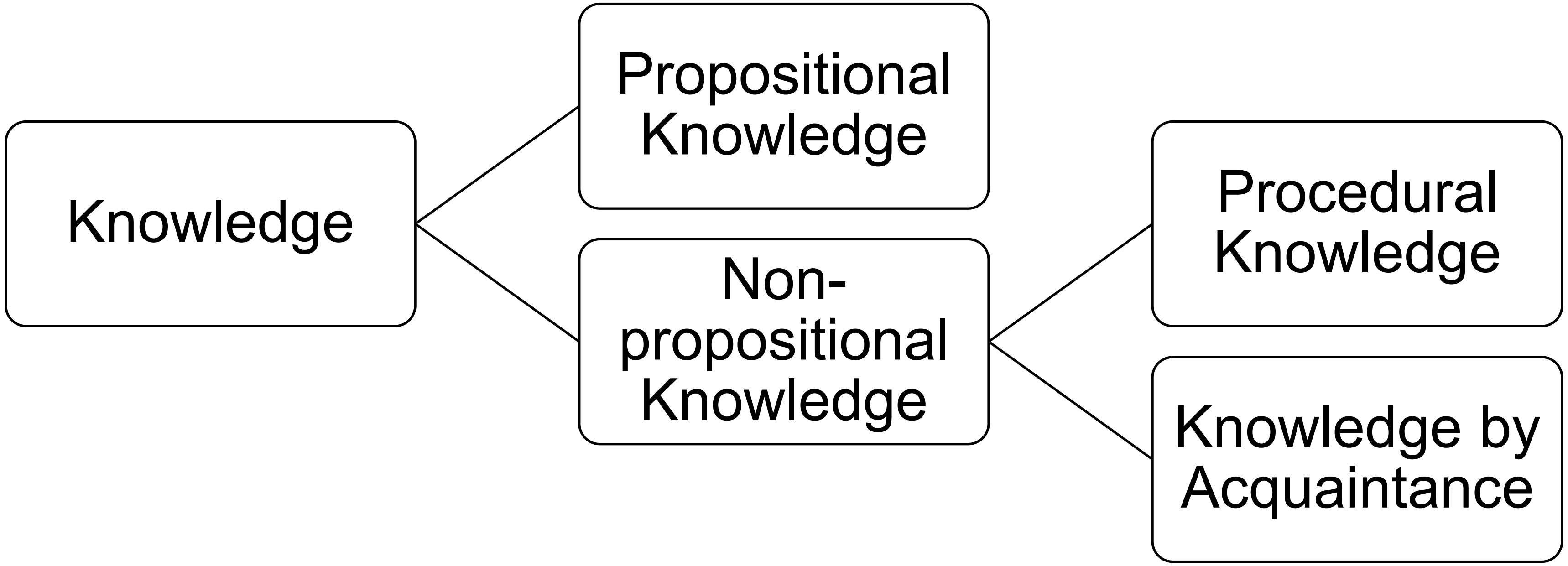


Agenda

- What is Procedural Knowledge?
- Tasks requiring Procedural knowledge.

What is Procedural Knowledge?

- Psychology View

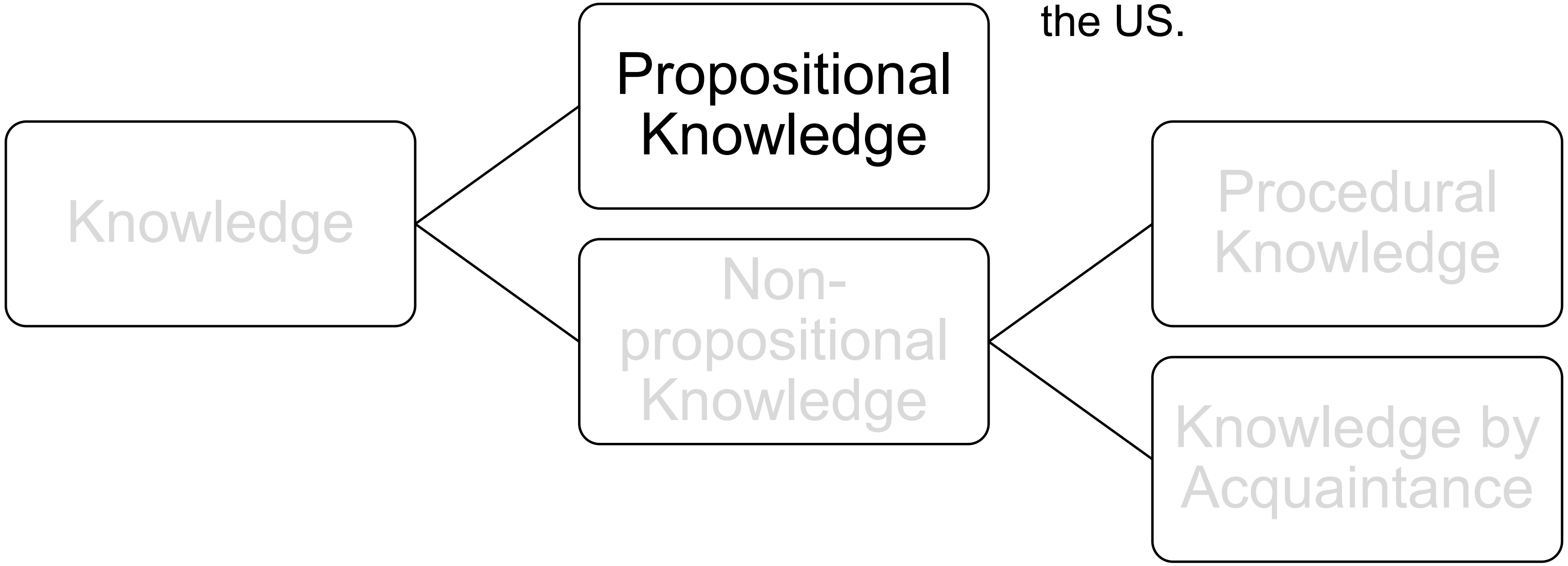


https://www.rep.routledge.com/articles/thematic/knowledge-concept-of/v-1/sections/the-varieties-of-knowledge

What is Procedural Knowledge?

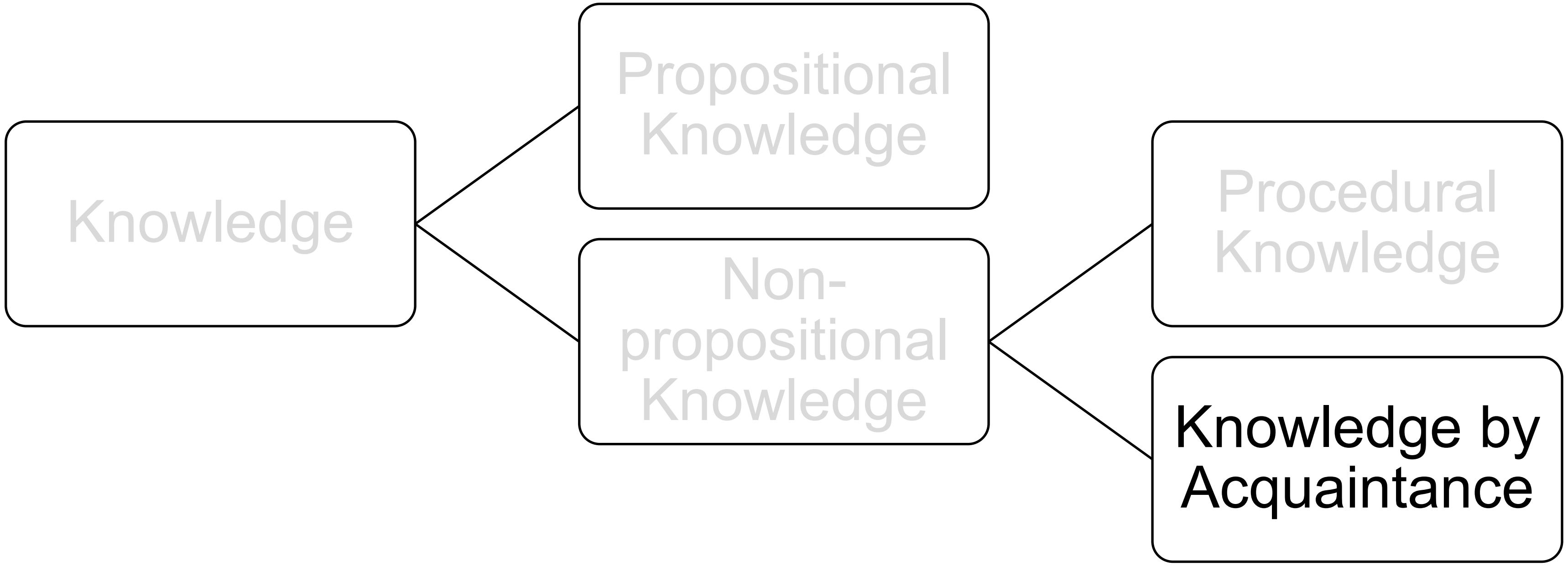
- Psychology View

I know Washington, D.C. is the capitol of the US.



What is Procedural Knowledge?

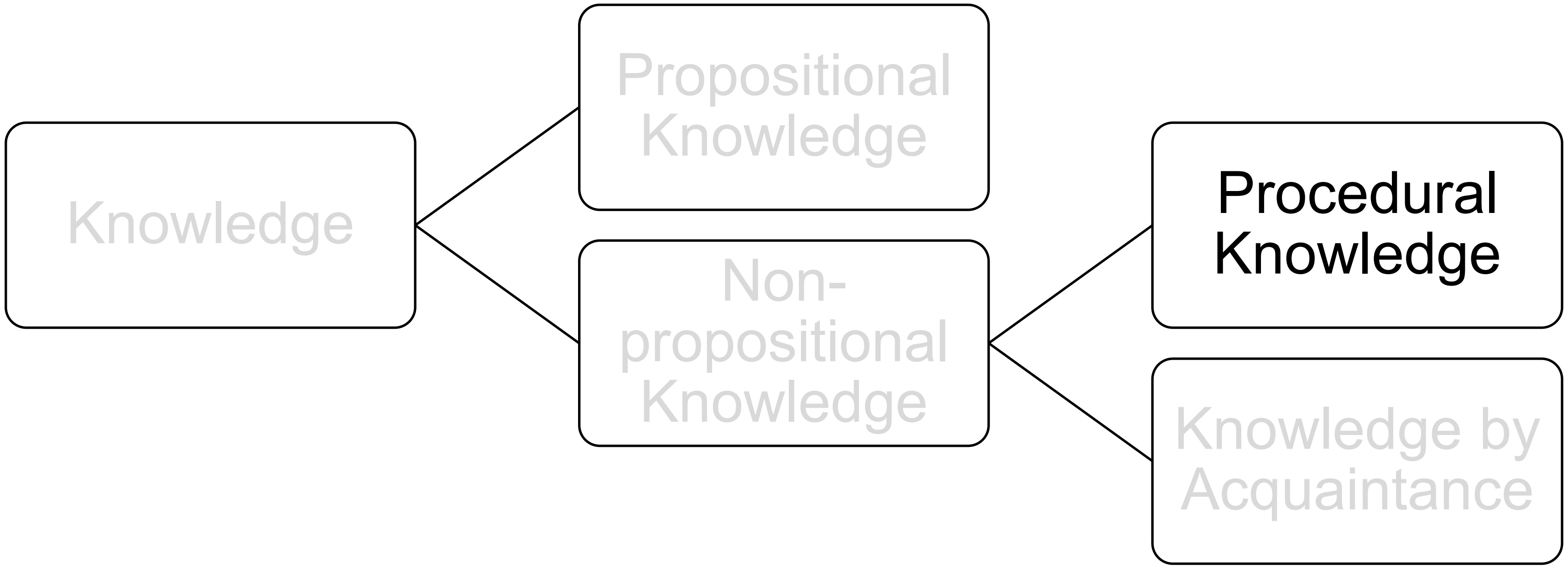
- Psychology View



I know someone.

What is Procedural Knowledge?

- Psychology View



I know **how** to do something.

Tasks Requiring Procedural Knowledge

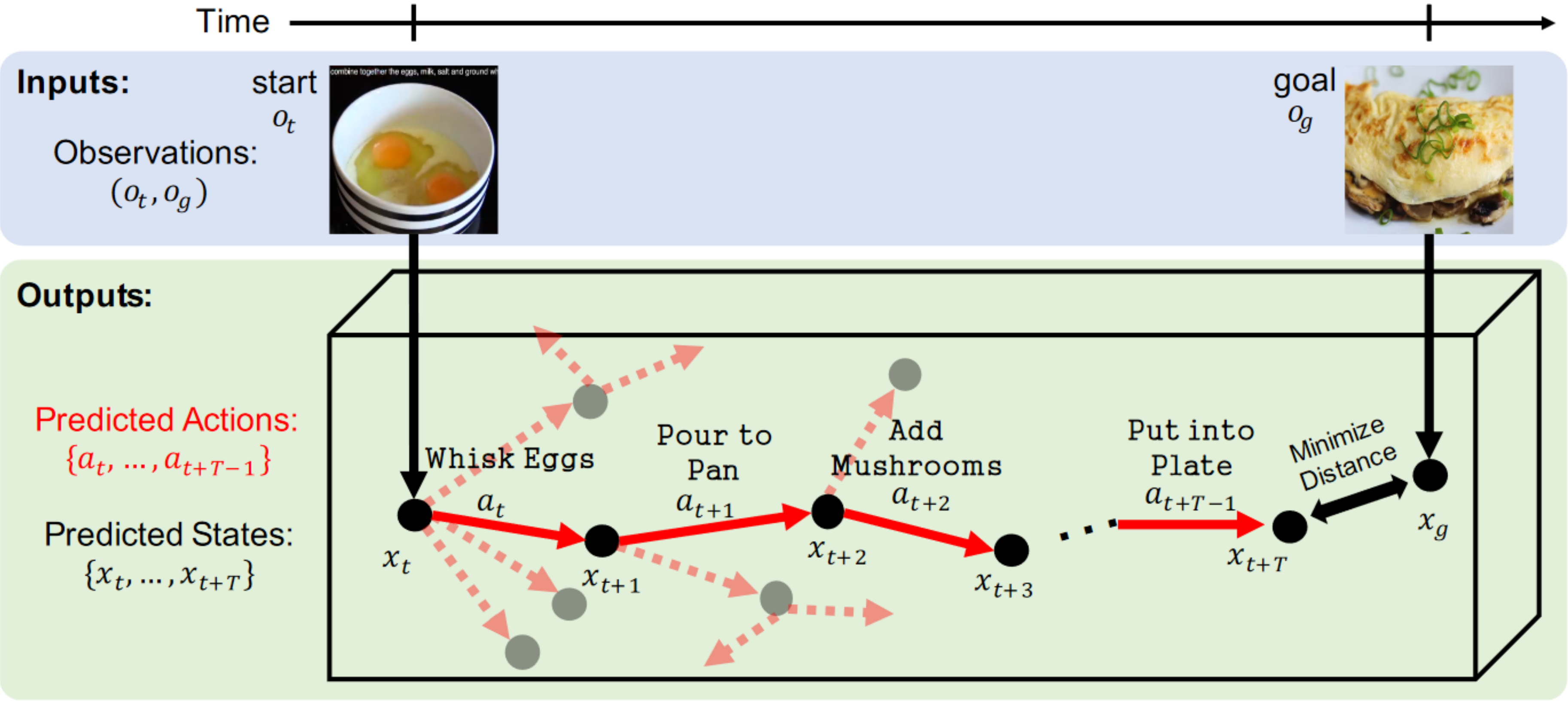
- Procedural planning

Given a start image and an end image, generate a sequence of actions.

Chang, Chien-Yi, et al. "Procedure planning in instructional videos." *European Conference on Computer Vision*. Springer, Cham, 2020.

Tasks Requiring Procedural Knowledge

- Procedural planning



Given a start image and an end image, generate a sequence of actions.

Chang, Chien-Yi, et al. "Procedure planning in instructional videos." *European Conference on Computer Vision*. Springer, Cham, 2020.

Tasks Requiring Procedural Knowledge

- Step forecasting

Time

What is the next step?

Given the historical video, predict the next step.

Frames are from Gordon Ramsay's **Fillet of Beef Wellington**

Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Tasks Requiring Procedural Knowledge

- Step forecasting

Time

What is the next step?

Assembling: Shingle the prosciutto on the plastic wrap; Spread mushroom over prosciutto; …

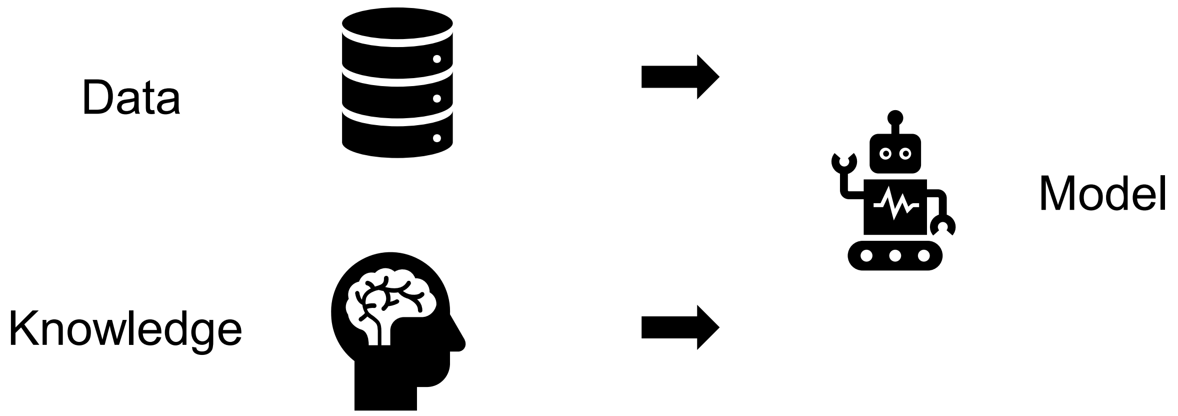
Given the historical video, predict the next step.

Frames are from Gordon Ramsay’s **Fillet of Beef Wellington**

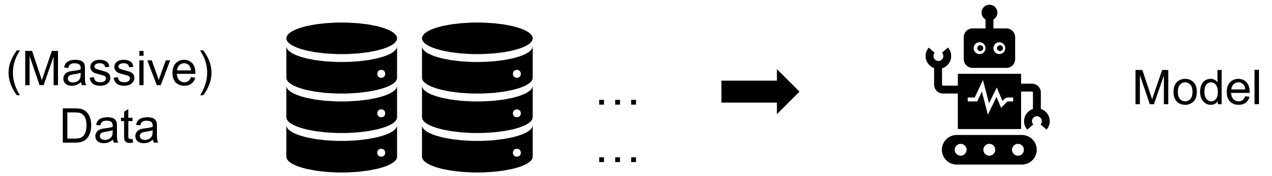
Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Agenda

- Explicit Knowledge Source: Learning with the help of external knowledge

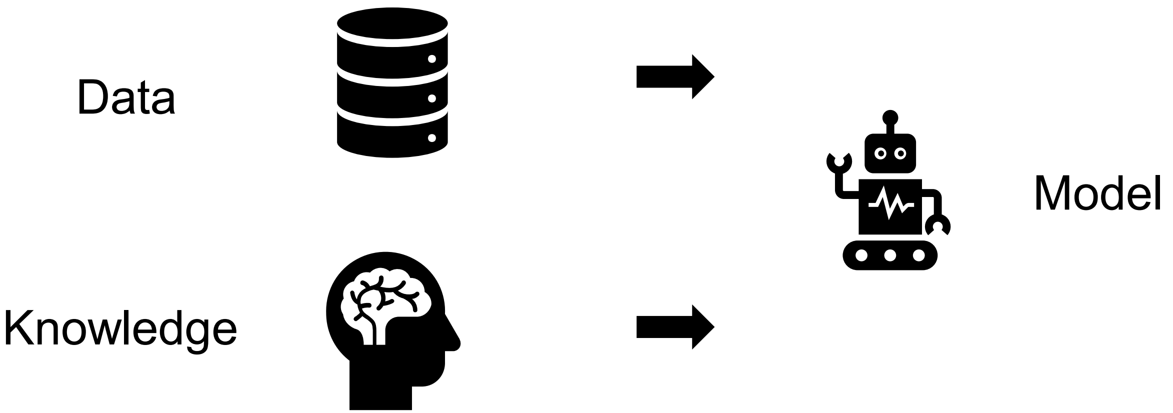


- Implicit Knowledge Source: Learning procedural knowledge from data



Agenda

- Explicit Knowledge Source: Learning with the help of external knowledge



Explicit Knowledge Source

- Procedural knowledge can be easily curated from the Internet
  - Recipe1M

Ingredients

- 3 lbs salmon
- 1 teaspoon cajun seasoning
- 1 tablespoon olive oil

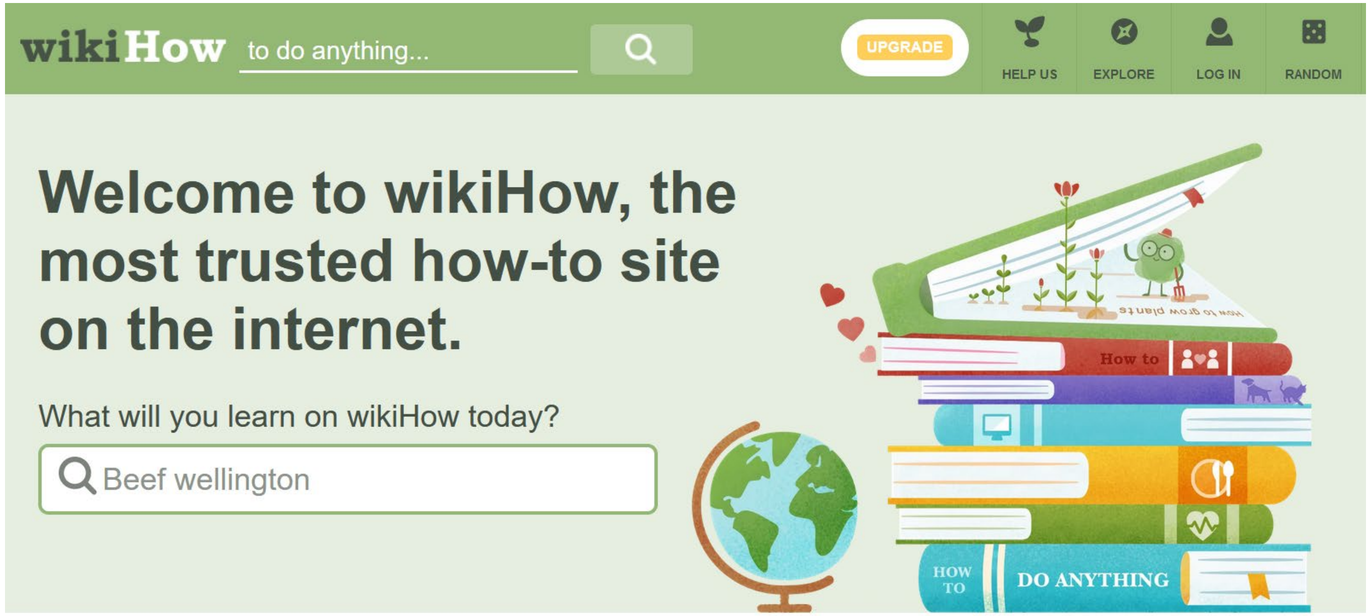
Cooking instructions

1. Rinse off salmon and pat dry with paper towel.
2. Drizzle cookie sheet with olive oil.
3. Place salmon (skin side down) on cookie sheet and drizzle more oil on top.
4. Shake Cajun seasoning on salmon to taste.
5. Broil 15-20 minutes or until center of salmon is done.

Salvador, Amaia, et al. "Learning cross-modal embeddings for cooking recipes and food images." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Explicit Knowledge Source

- Procedural knowledge can be easily curated from the Internet
  - Recipe1M
  - wikiHow



Step 1. Sear the filiet mignon to brown.

Over high heat, coat bottom of a heavy skillet with olive oil. Once pan is nearly smoking, sear tenderloin until well-browned on all sides.

Step 2. Fry the mushroom until they are dried.

To skillet, add butter and melt over medium heat. Add mushroom mixture and cook until liquid has evaporated.

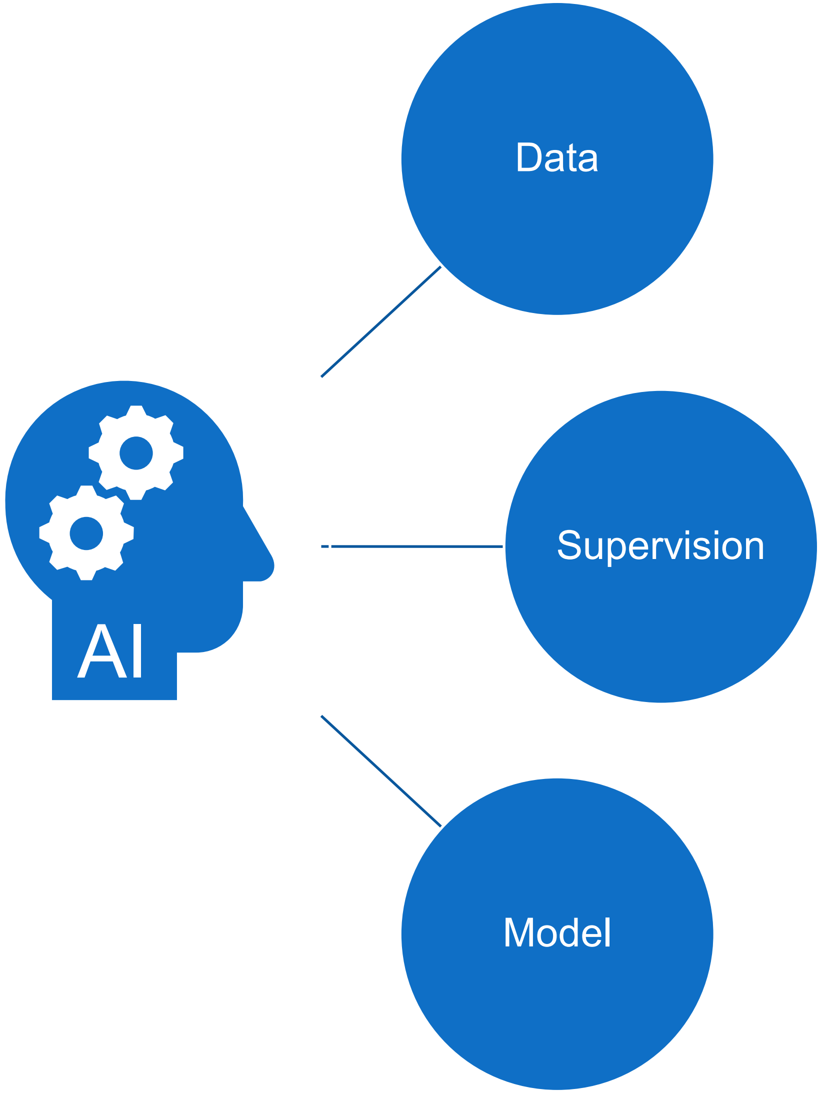
Step 3. Assembling.

Shingle the prosciutto on the plastic wrap into a rectangle that's big enough to cover the whole tenderloin. Spread the duxelles evenly and thinly over the prosciutto.

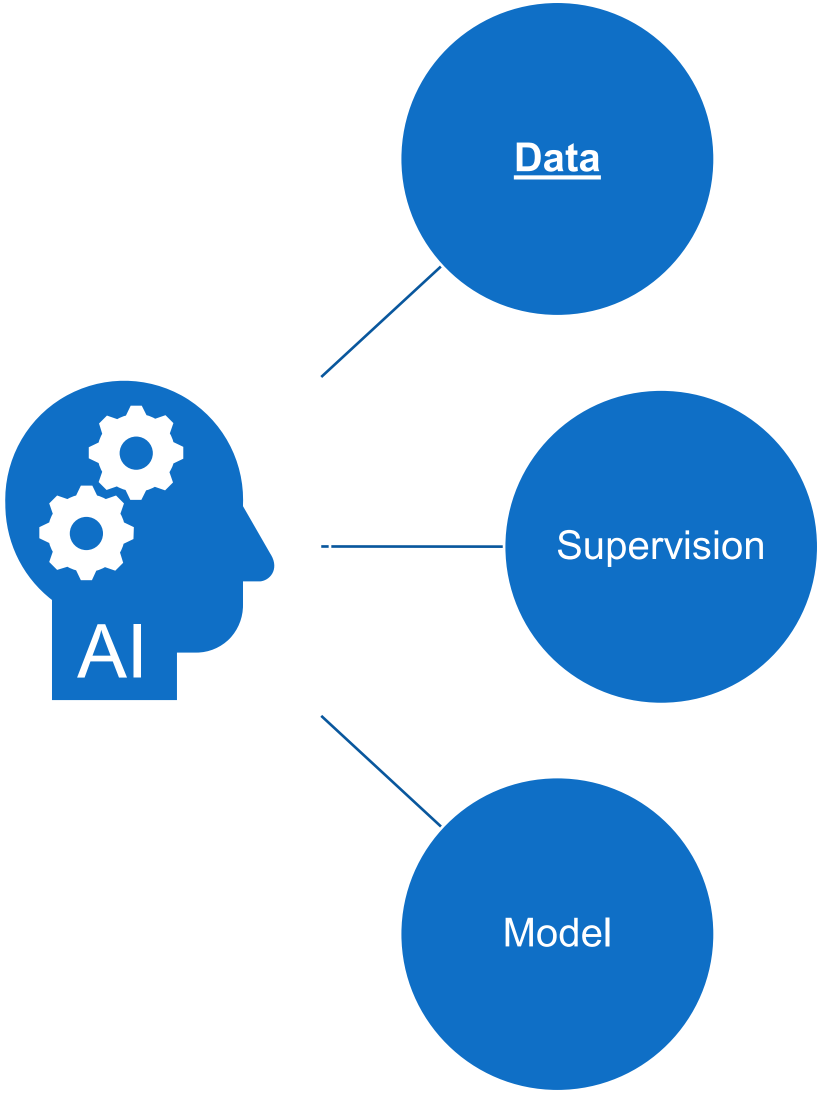
……

https://www.wikihow.com/Main-Page
Koupaee, Mahnaz, and William Yang Wang. "Wikihow: A large scale text summarization dataset." *arXiv preprint arXiv:1810.09305* (2018).

How to Utilize the Knowledge Source?

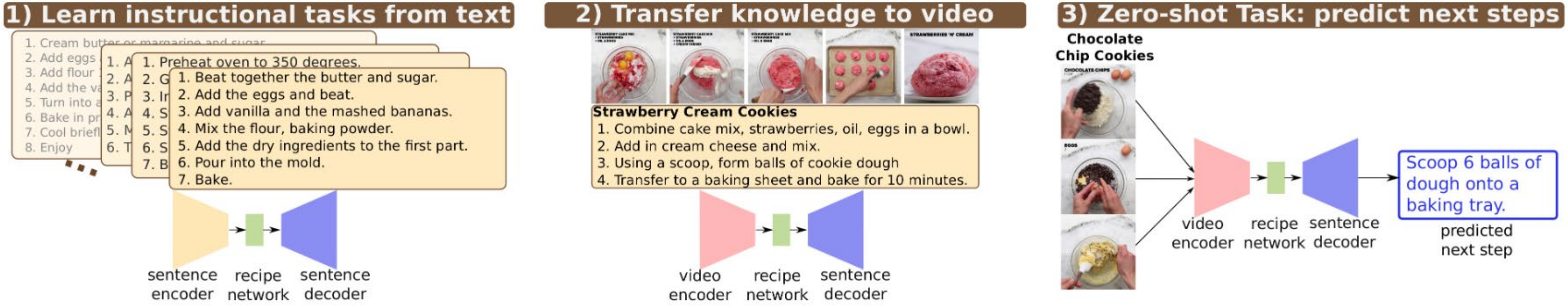


How to Utilize the Knowledge Source?



Zero-Shot Anticipation for Instructional Activities

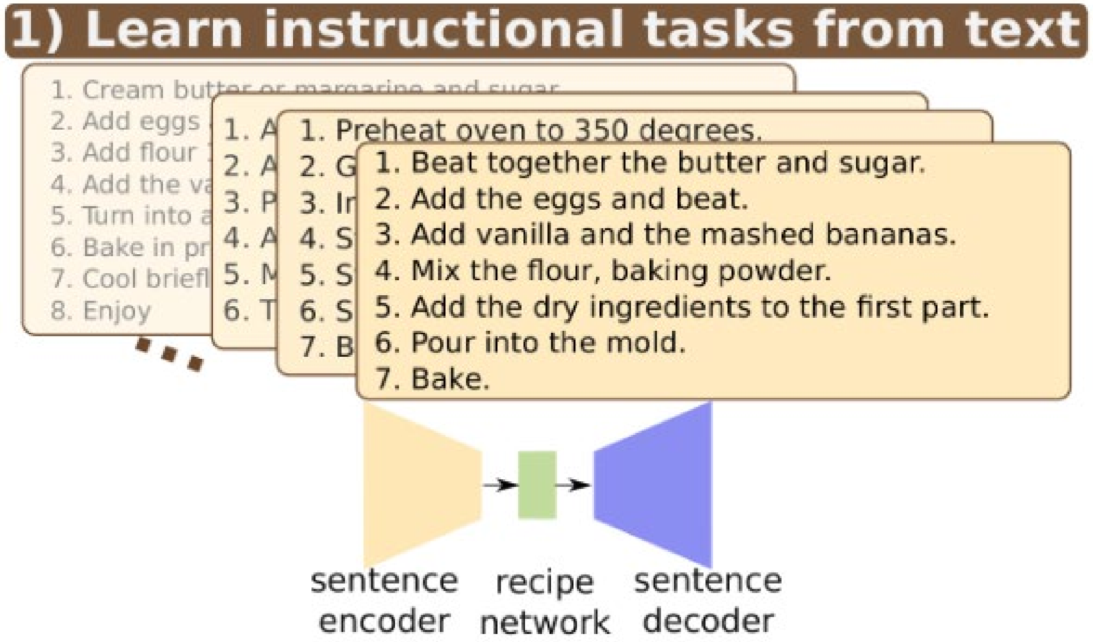
- Key Idea: Obtain training data from knowledge base.

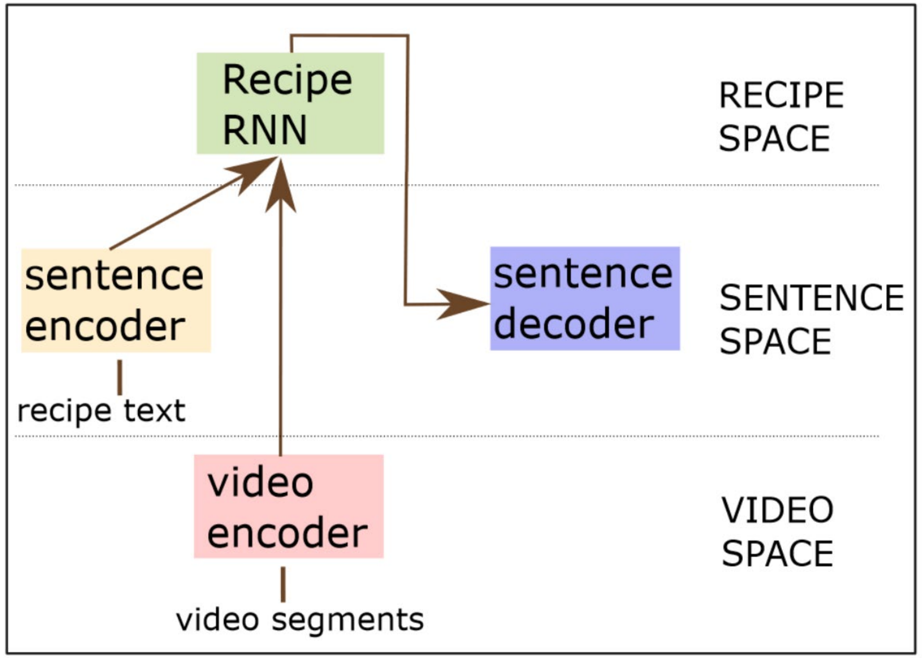


Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Sener, Fadime, Rishabh Saraf, and Angela Yao. "Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Zero-Shot Anticipation for Instructional Activities

- Sentence encoder encodes a step sentence into a step vector.
- Recipe network is a RNN modeling procedures.
- Sentence decoder decodes step sentences.



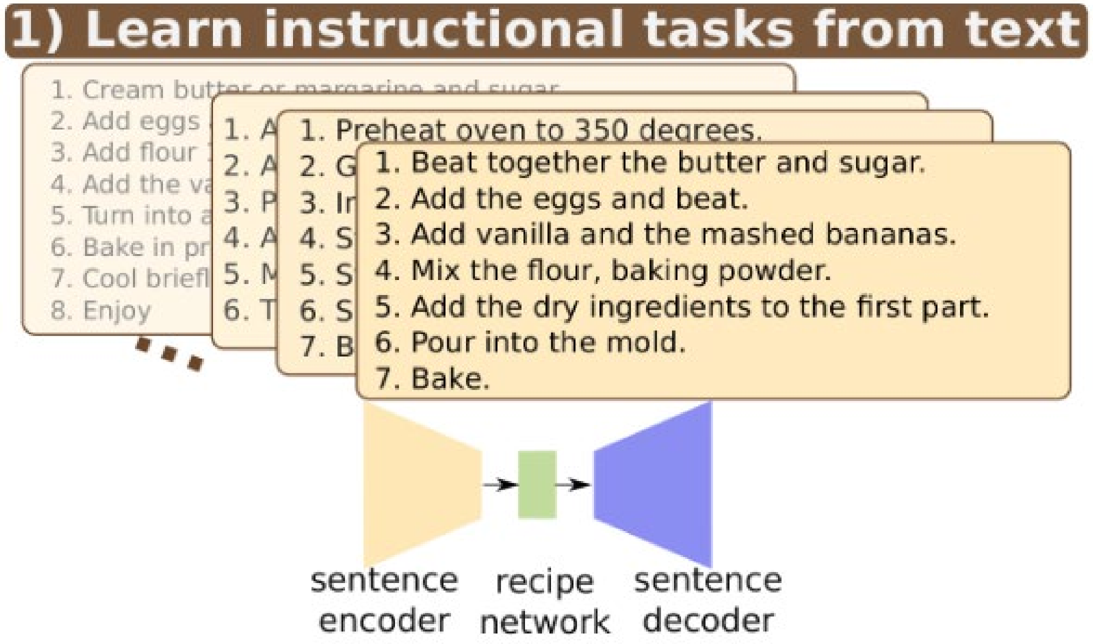


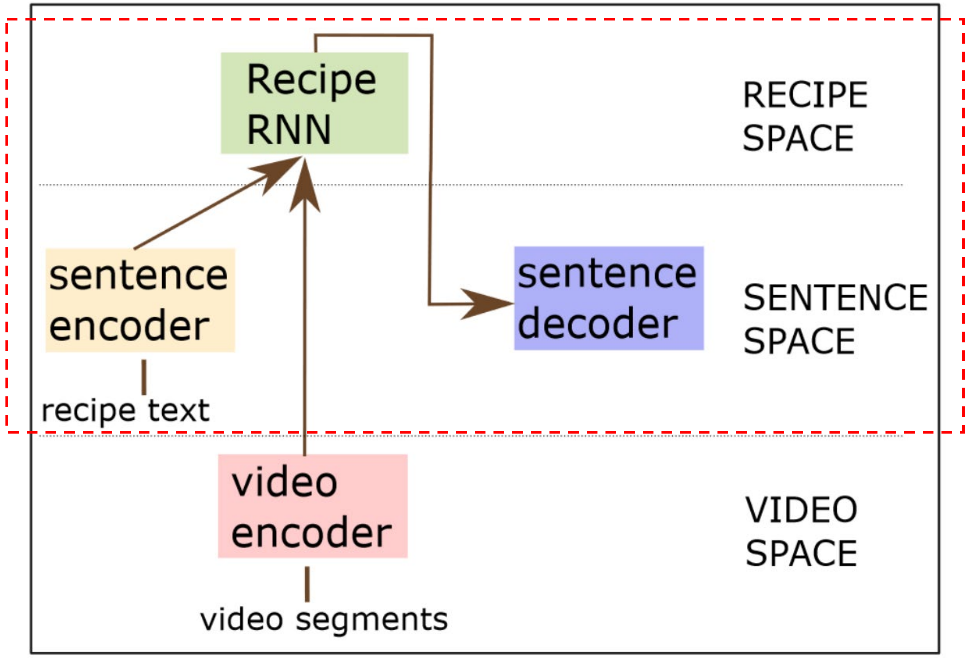
Model Overview

Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Sener, Fadime, Rishabh Saraf, and Angela Yao. "Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Zero-Shot Anticipation for Instructional Activities

- Sentence encoder encodes a step sentence into a step vector.
- Recipe network is a RNN modeling procedures.
- Sentence decoder decodes step sentences.



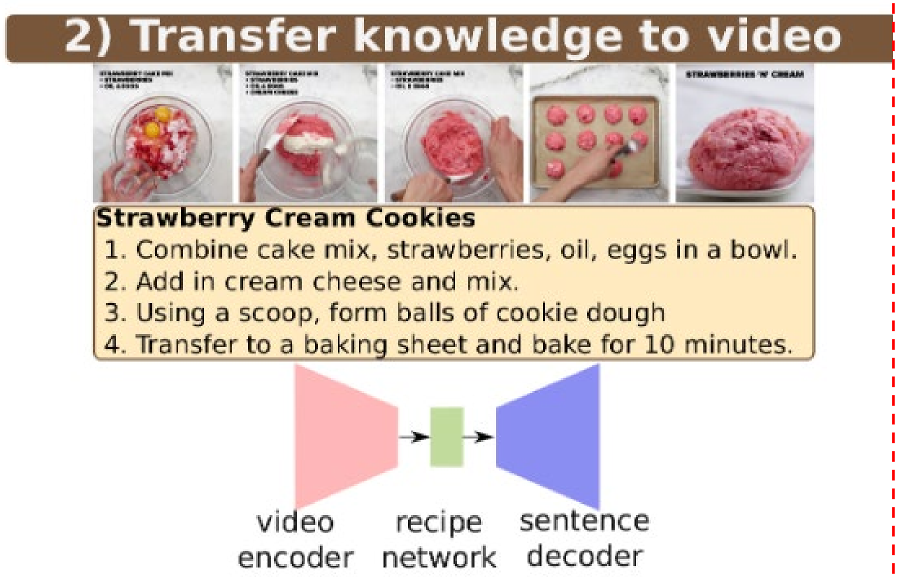


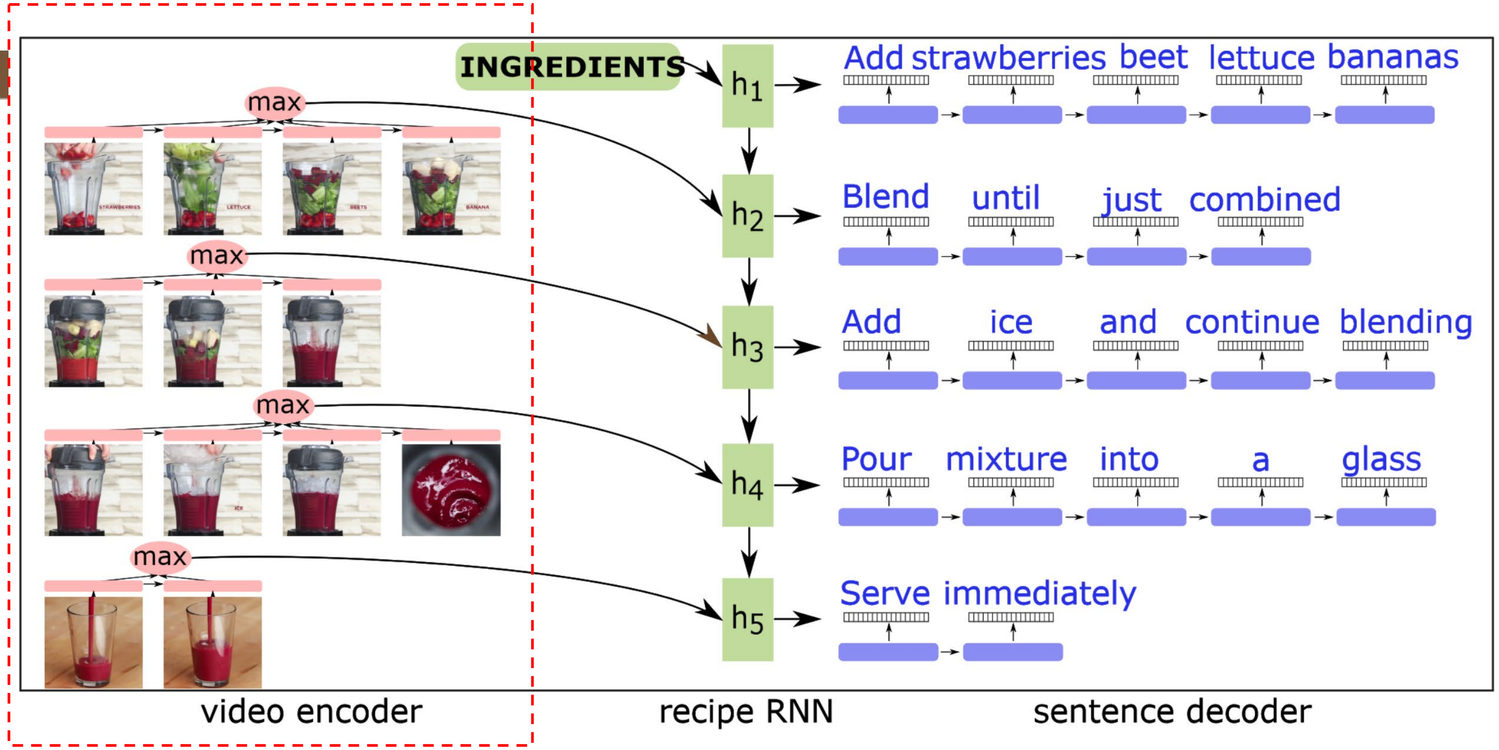
Model Overview

Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Sener, Fadime, Rishabh Saraf, and Angela Yao. "Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Zero-Shot Anticipation for Instructional Activities

- Only train the video encoder to project video into step vectors with annotated data.

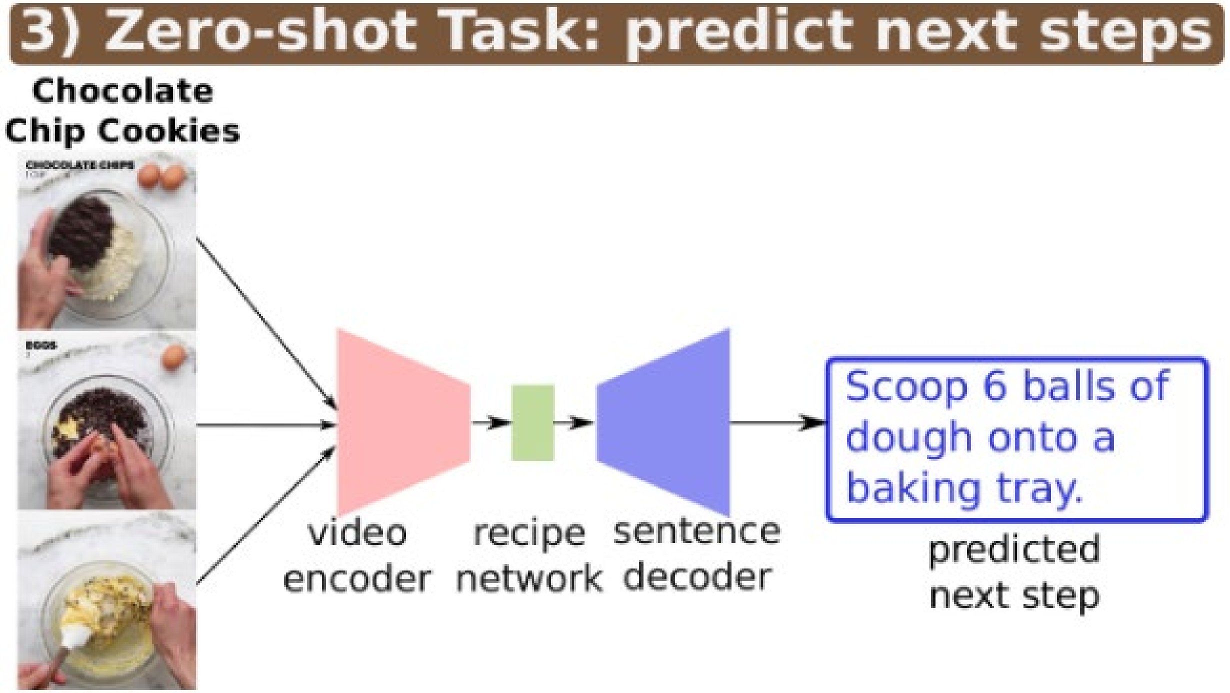




Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Sener, Fadime, Rishabh Saraf, and Angela Yao. "Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Zero-Shot Anticipation for Instructional Activities

- Generalize on new tasks.



Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
Sener, Fadime, Rishabh Saraf, and Angela Yao. "Transferring Knowledge from Text to Video: Zero-Shot Anticipation for Procedural Actions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Zero-Shot Anticipation for Instructional Activities

- Strong zeros-hot performance on the proposed Tasty video dataset

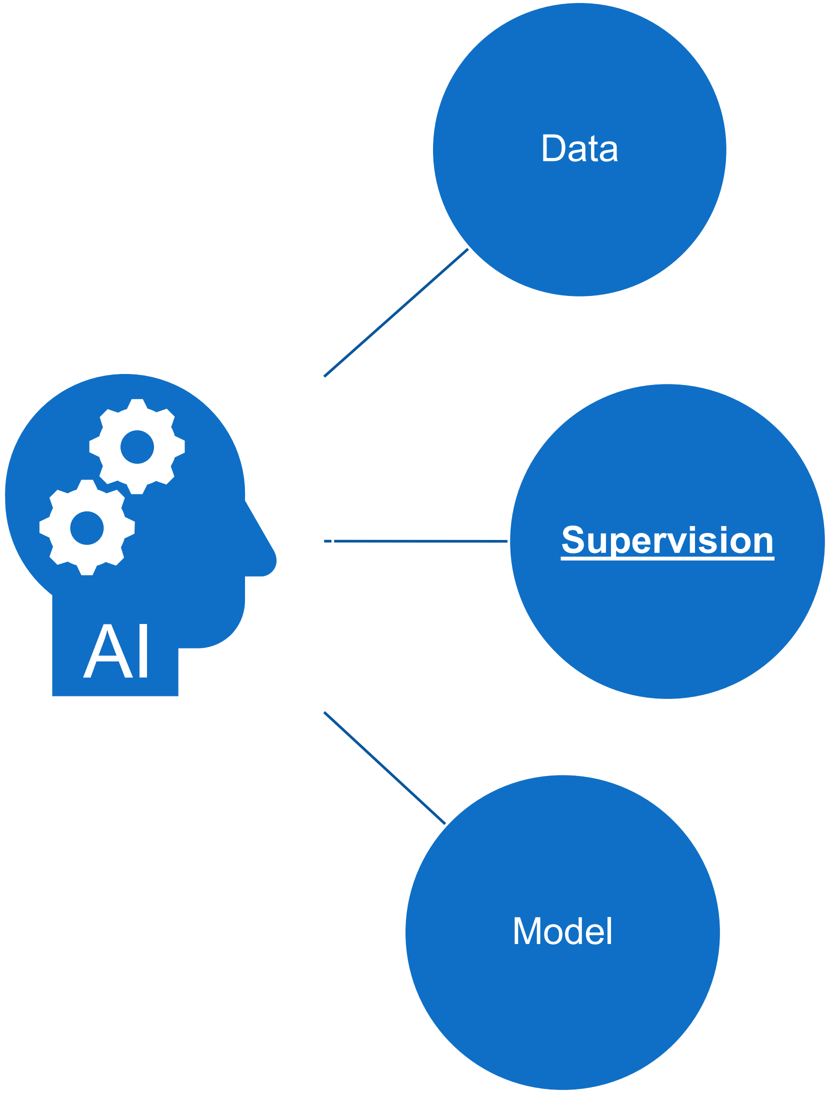
The larger knowledge base used, the better!

Method	ING	VERBS	BLEU1	BLEU4	METEOR
S2VT [53] (GT)	7.59	19.18	18.03	1.10	9.12
S2VT [53], next (GT)	1.54	10.66	9.14	0.26	5.59
End-to-end [60]	-	-	-	0.54	5.48
Ours Visual (GT)	20.40	19.18	19.05	1.48	11.78
Ours Visual	16.66	17.08	17.59	1.23	11.00
Ours Text (100%)	26.09	27.19	26.78	3.30	17.97
Ours Text (50%)	23.01	24.90	25.05	2.42	16.98
Ours Text (25%)	19.43	23.83	23.54	2.03	16.05
Ours Text (0%)	5.80	9.42	10.58	0.24	6.80
Ours Text noING	9.04	22.00	20.11	0.92	13.07
Ours joint video-text	22.27	23.35	21.75	2.33	14.09

- Limitation
  - Domain is limited to cooking.
  - Rely on annotated data samples for training video encoder.

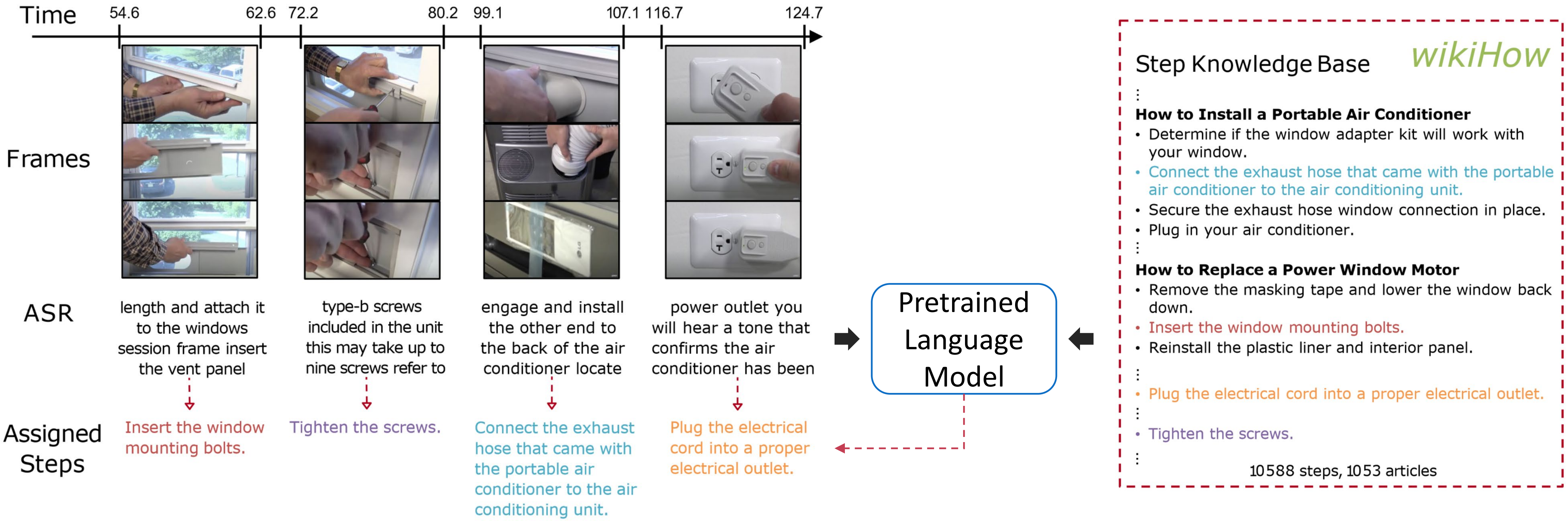
Sener, Fadime, and Angela Yao. "Zero-shot anticipation for instructional activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

How to Utilize the Knowledge Source?



Learning To Recognize Procedural Activities with Distant Supervision

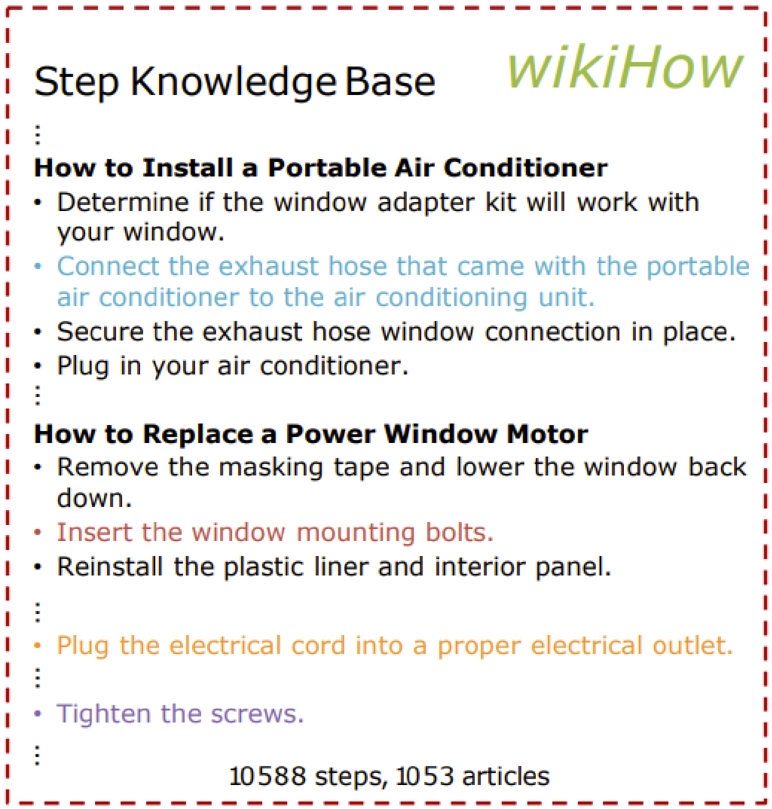
- Key Idea: Leverage pretrained language model to align knowledge base and videos with speech to obtain supervision.



Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

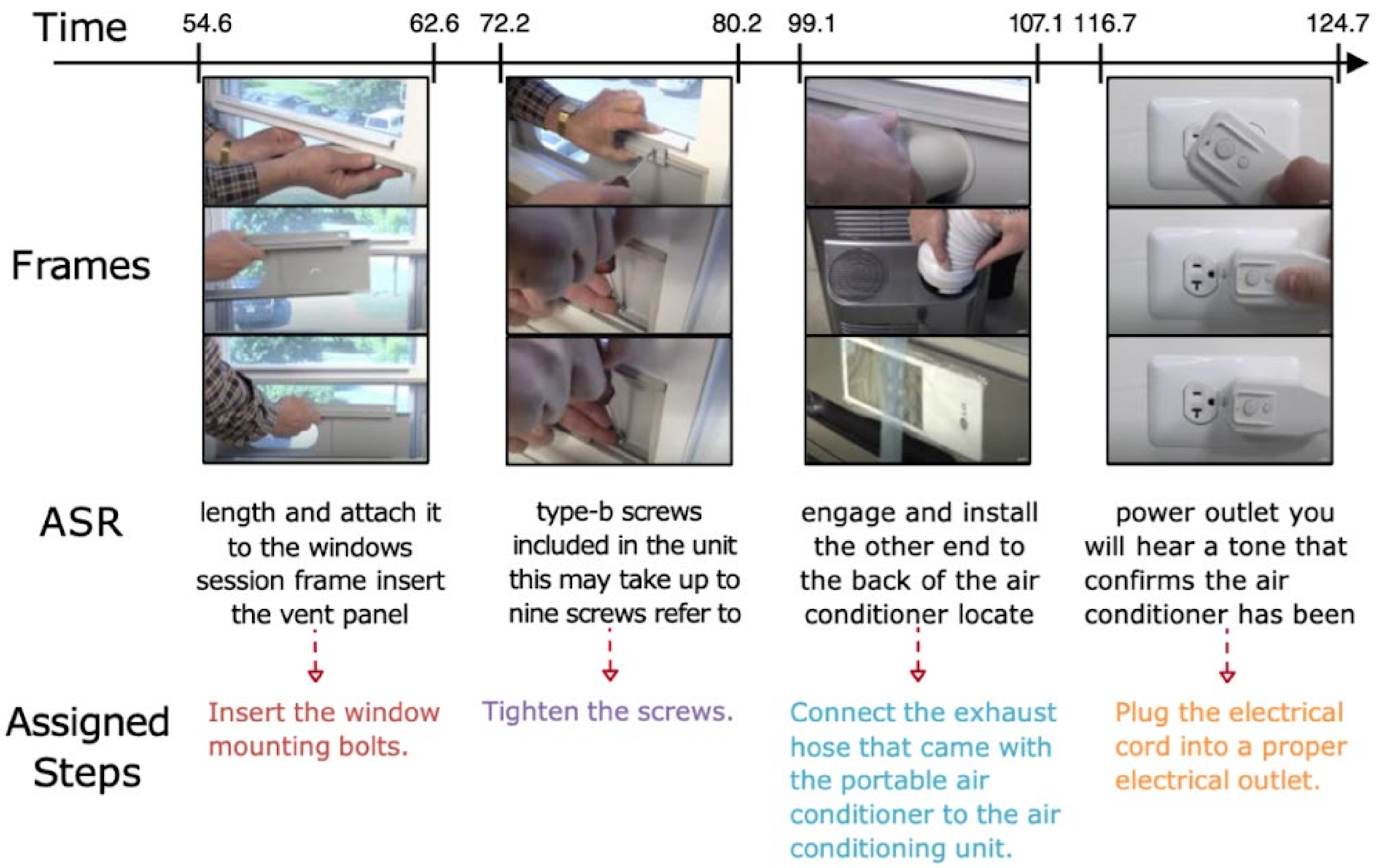
Learning To Recognize Procedural Activities with Distant Supervision

- Step Knowledge Base Construction
    - Use 1053 tasks, each of which has at least 100 examples in the HowTo100M dataset
    - Find the correspond articles on WikiHow
    - Collect sentences for each step in each of the tasks



Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Learning To Recognize Procedural Activities with Distant Supervision

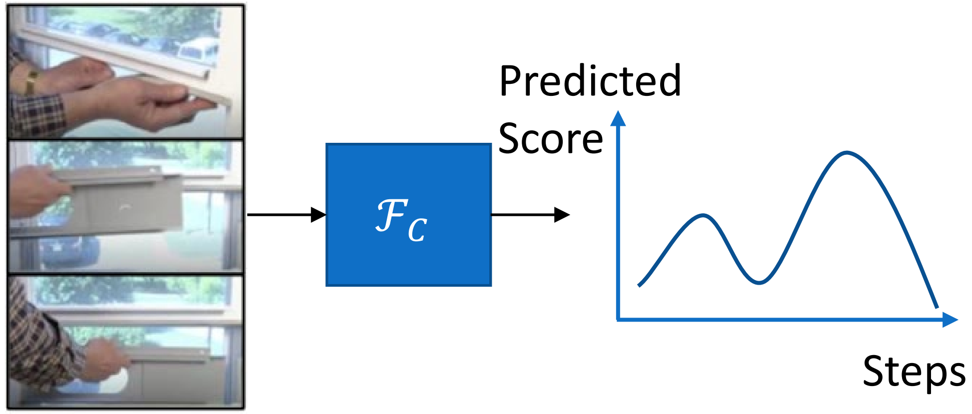


- Distant supervision creation
  - Leverage a pretrained language model to produce embeddings for both **steps** and **ASR sentences** from the video.
  - Then calculate similarity between each ASR sentence and all the steps.

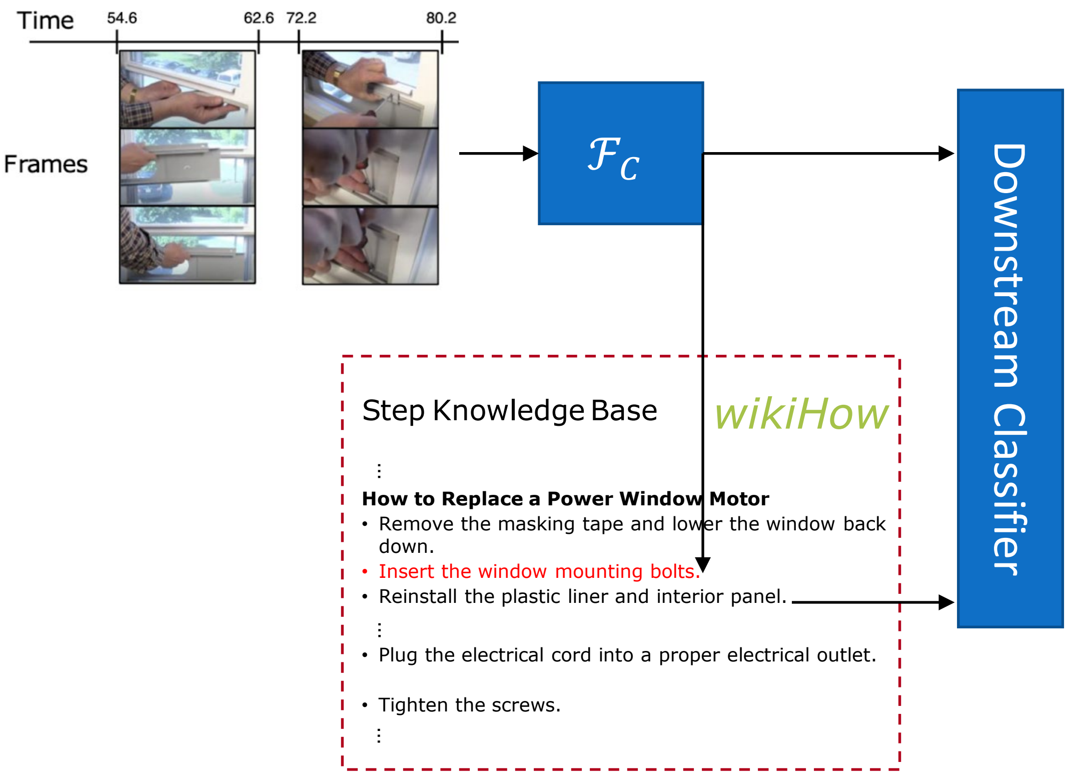
Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Learning To Recognize Procedural Activities with Distant Supervision

<u>Pretraining: Learning to align videos and the step knowledge base</u>



<u>Finetuning: Training a classifier with both step-level video representation and ordering information from the knowledge base</u>



Lin, Xudong, et al. "Learning to recognize procedural activities with distant supervision." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Learning To Recognize Procedural Activities with Distant Supervision

- Step Forecasting on COIN
  - Wikihow Knowledge provides high-quality distant supervision!
  - Ordering information in the knowledge base further helps!

Long-term Model	Segment Model	Pretraining Supervision	Pretraining Dataset	Acc (%)
Basic Transformer	S3D [39]	Unsupervised: MIL-NCE on ASR	HT100M	28.1
Basic Transformer	SlowFast [17]	Supervised: action labels	Kinetics	25.6
Basic Transformer	TimeSformer [8]	Supervised: action labels	Kinetics	34.7
Basic Transformer	TimeSformer [8]	Unsupervised: k-means on ASR	HT100M	34.0
Basic Transformer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**38.2**
Transformer w/ KB Transfer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**39.4**

- The supervision from the wikihow knowledge base also helps

Recognition of procedural activities on COIN

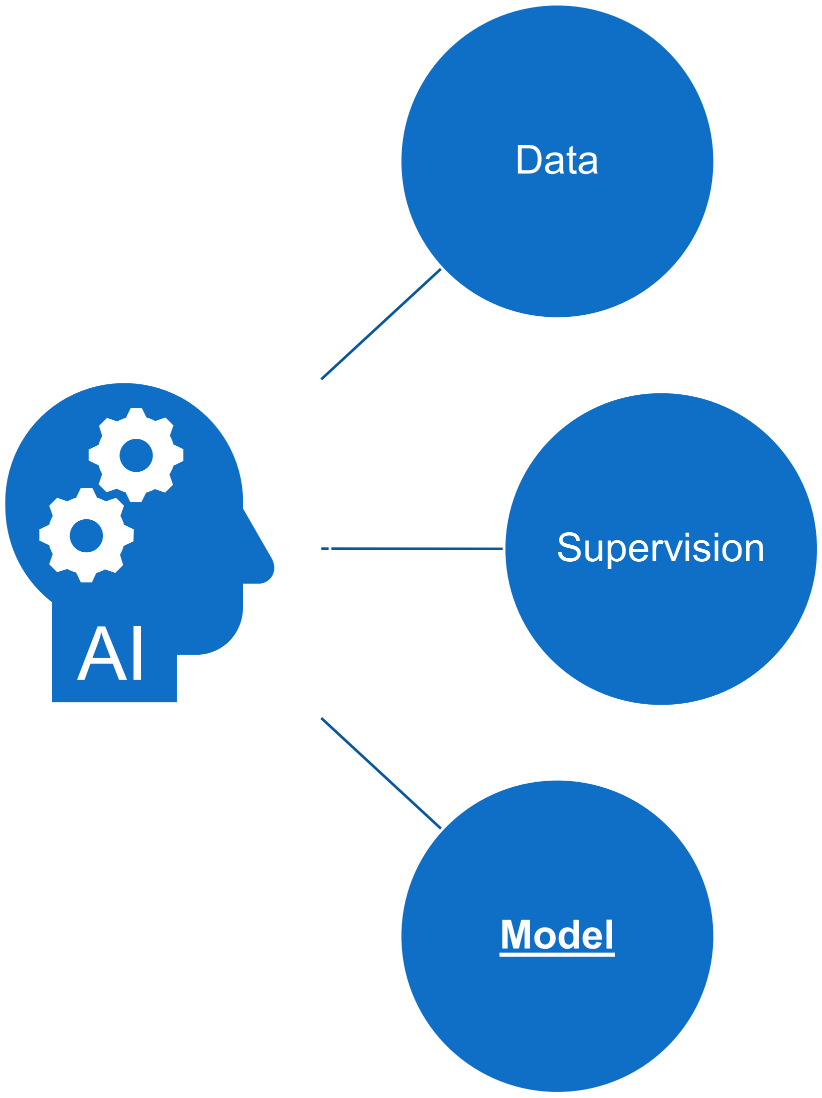
Long-term Model	Segment Model	Pretraining Supervision	Pretraining Dataset	Acc (%)
TSN (RGB+Flow) [57]	Inception [54]	Supervised: action labels	Kinetics	73.4*
Basic Transformer	S3D [39]	Unsupervised: MIL-NCE on ASR	HT100M	70.2*
Basic Transformer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**88.9**
Transformer w/ KB Transfer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**90.0**

Egocentric video classification

Segment Model	Pretraining Supervision	Pretraining Dataset	Action (%)	Verb (%)	Noun (%)
ViViT-L [6]	Supervised: action labels	Kinetics	44.0	66.4	56.8
TimeSformer [8]	Supervised: action labels	Kinetics	42.3	66.6	54.4
TimeSformer	**Unsupervised: distant supervision (ours)**	HT100M	44.4	67.1	**58.1**

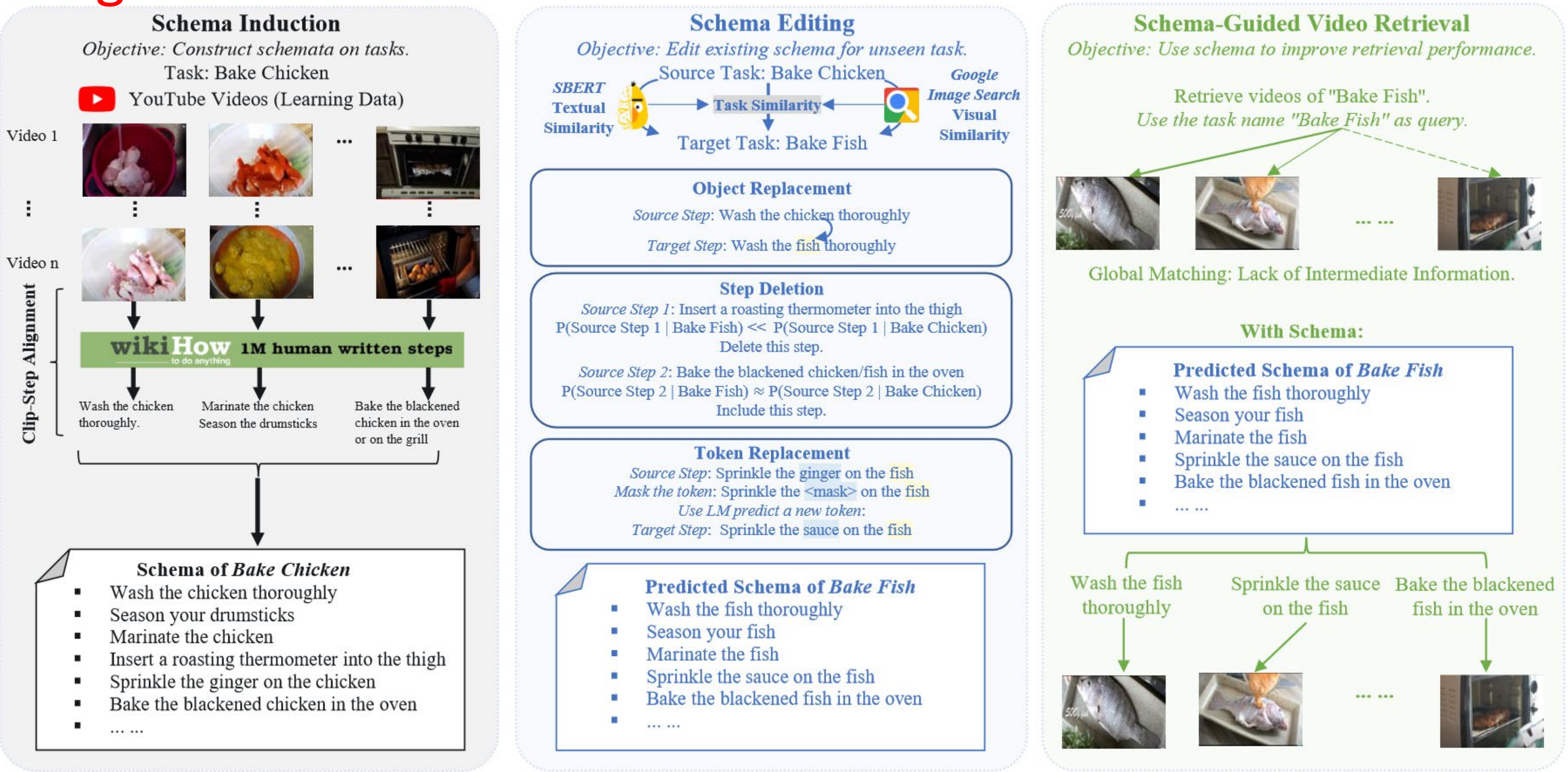
- Limitation: Didn't employ ordering information in the pretraining model.

How to Utilize the Knowledge Source?



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

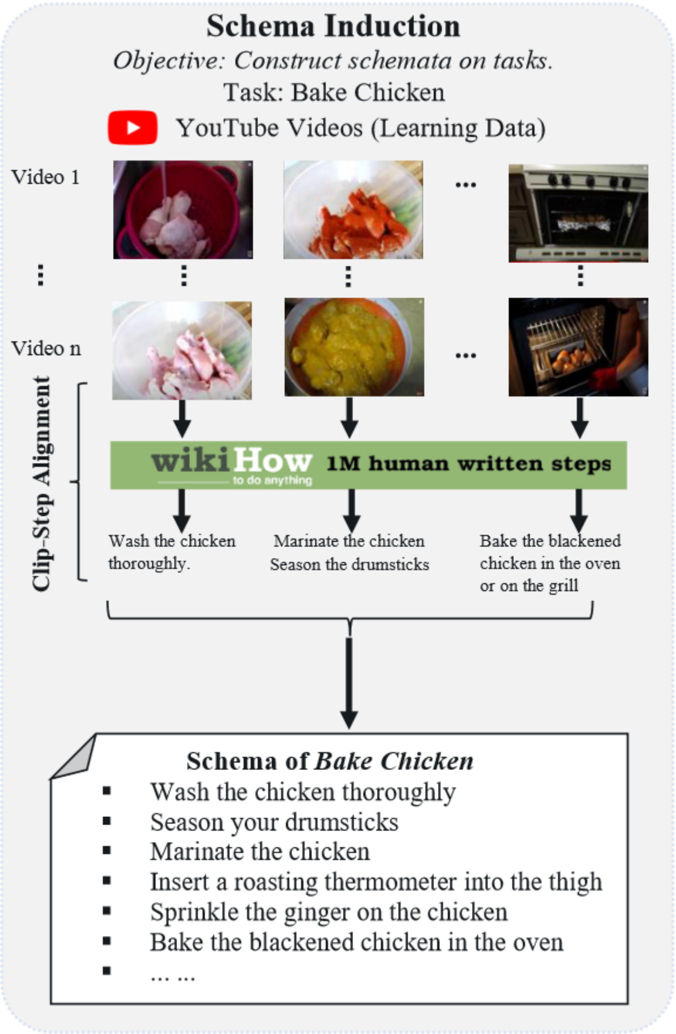
- Key Idea: Learning multimodal schema to represent procedural knowledge.



Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

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

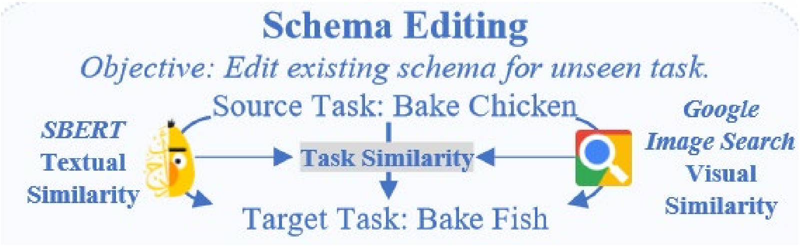
- Schema Induction
  - For each task, find corresponding steps from wikiHow and videos from YouTube.
  - For each segment in each video, retrieve most relevant steps with existing video-text matching models.



Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

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

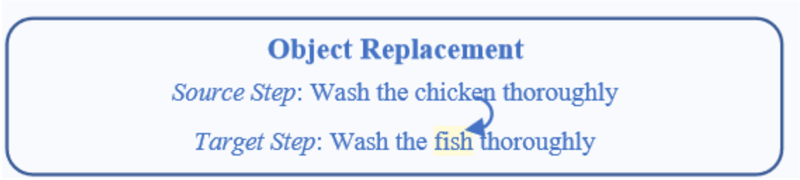
- Schema Editing
  - For an unseen task, find the most similar seen task based on both textual and visual similarity.



Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

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

- Schema Editing
  - For an unseen task, find the most similar seen task based on both textual and visual similarity
  - Replace object towards the unseen task.



Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

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

- Schema Editing
  - For an unseen task, find the most similar seen task based on both textual and visual similarity
  - Replace object towards the unseen task.
  - Delete steps that are not relevant in the new task with a pretrained language model.

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.

Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

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

- Schema Editing
  - For an unseen task, find the most similar seen task based on both textual and visual similarity
  - Replace object towards the unseen task.
  - Delete steps that are not relevant in the new task with a pretrained language model.
  - Replace tokens least likely associated with the task in each step by prompting a pretrained language model.

Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

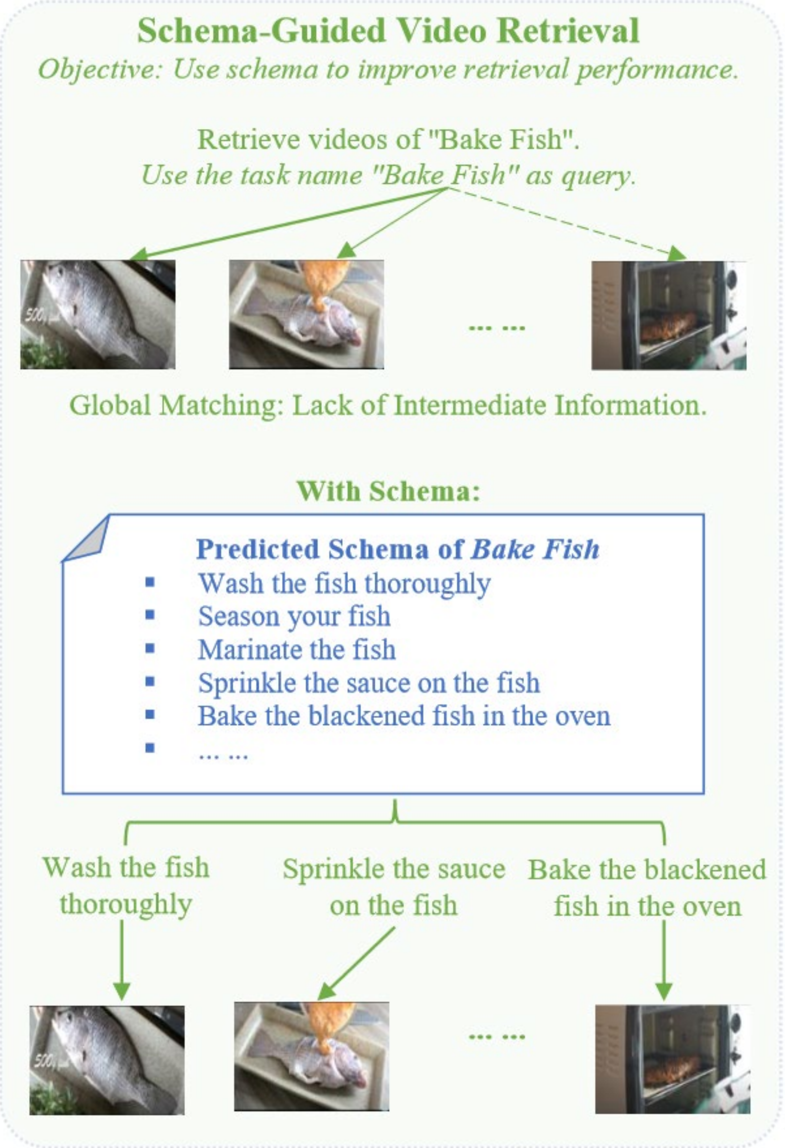
Induce, Edit, Retrieve: Language Grounded Multimodal Schema for Instructional Video Retrieval

- The learned schema provides step-level information to better retrieve videos.

	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

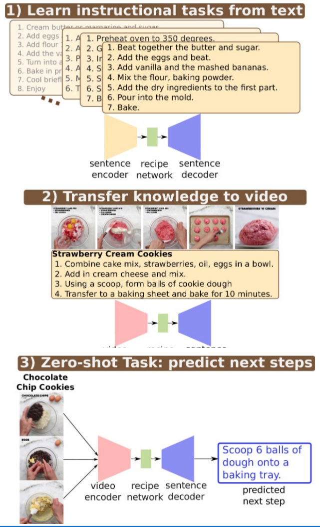
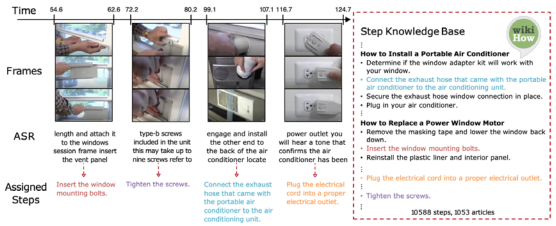
<u>Even comparable with oracle (using manual step annotation for each query)</u>

- Limitation
  - Schema is restricted to step sequence without considering graph structures, e.g., optional/exchangeable steps.
  - Only evaluated on text-video retrieval.



Yang, Yue, et al. "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval." *arXiv preprint arXiv:2111.09276* (2021).

Summary of Methods Using Explicit Knowledge

	Sener & Yao ICCV 2019	Lin et al. CVPR 2022	Yang et al.
			
Knowledge as data	✓		
Knowledge as supervision		✓	
Knowledge for model		✓	✓
Sequential knowledge	✓	✓	✓
Multimodal knowledge			✓

Summary of Methods Using Explicit Knowledge

	Sener & Yao ICCV 2019	Lin et al. CVPR 2022	Yang et al.
Knowledge as data	✓		
Knowledge as supervision		✓	
Knowledge for model		✓	✓
Sequential knowledge	✓	✓	✓
Multimodal knowledge			✓

- What is next?

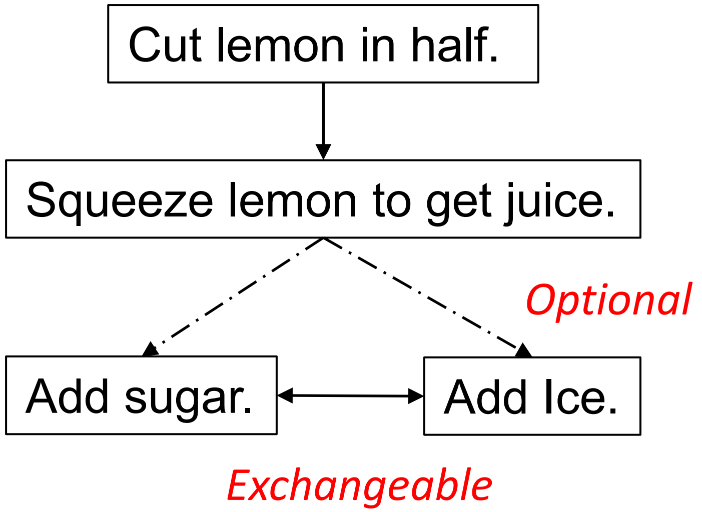
Future Challenge: Is sequential knowledge enough?

- Procedural knowledge:
  - From a sequence to a graph!

How to make lemonade?

1. Cut lemon in half.
2. Squeeze lemon to get juice.
3. Add sugar.
4. Add Ice.

Current Knowledge

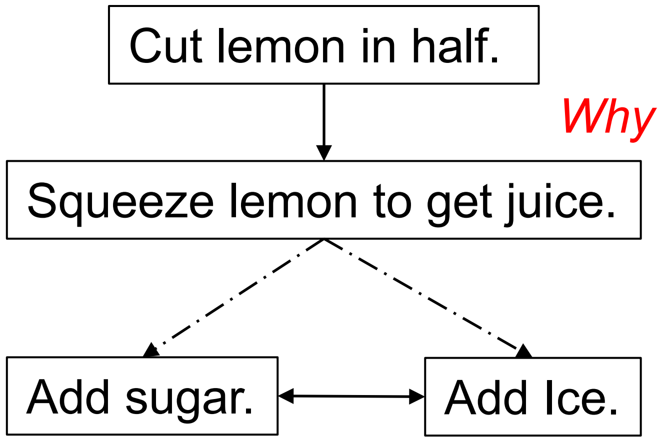


Reality

Future Challenge: Interpret but Not Memorize

- Do models understand **why** the steps are ordered as in the knowledge base?

How to make lemonade?

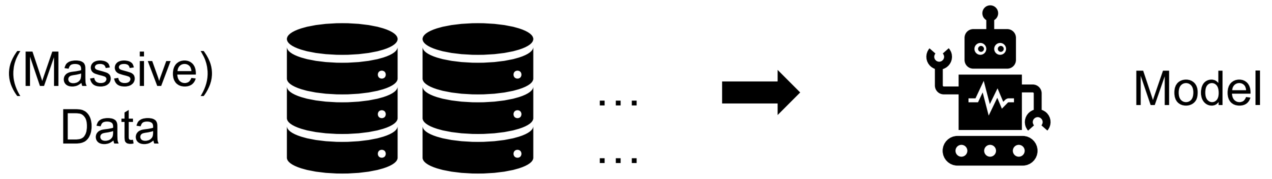


Why these two steps cannot be exchanged?

What is the intent of this step?

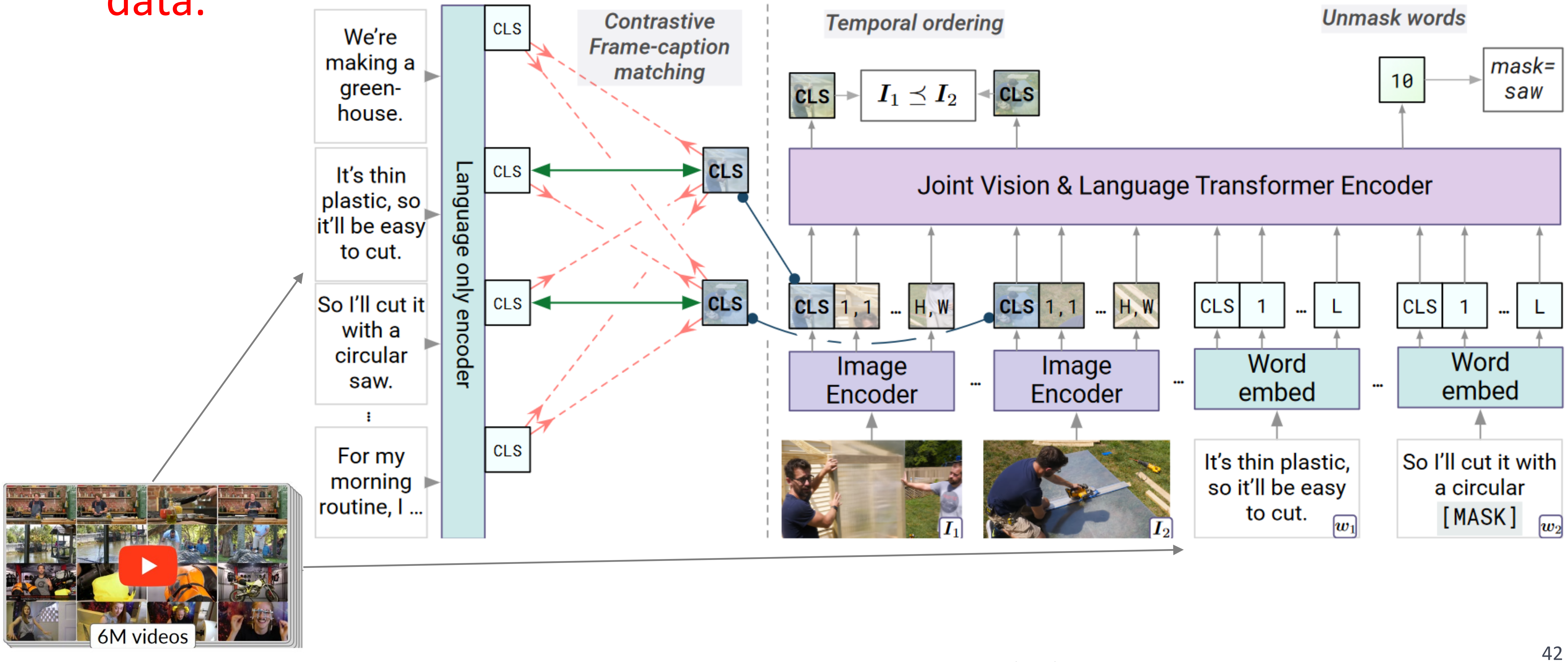
Agenda

- Explicit Knowledge Source: Learning with the help of external knowledge
- Implicit Knowledge Source: Learning procedural knowledge from data



MERLOT:
Multimodal Neural Script Knowledge Models

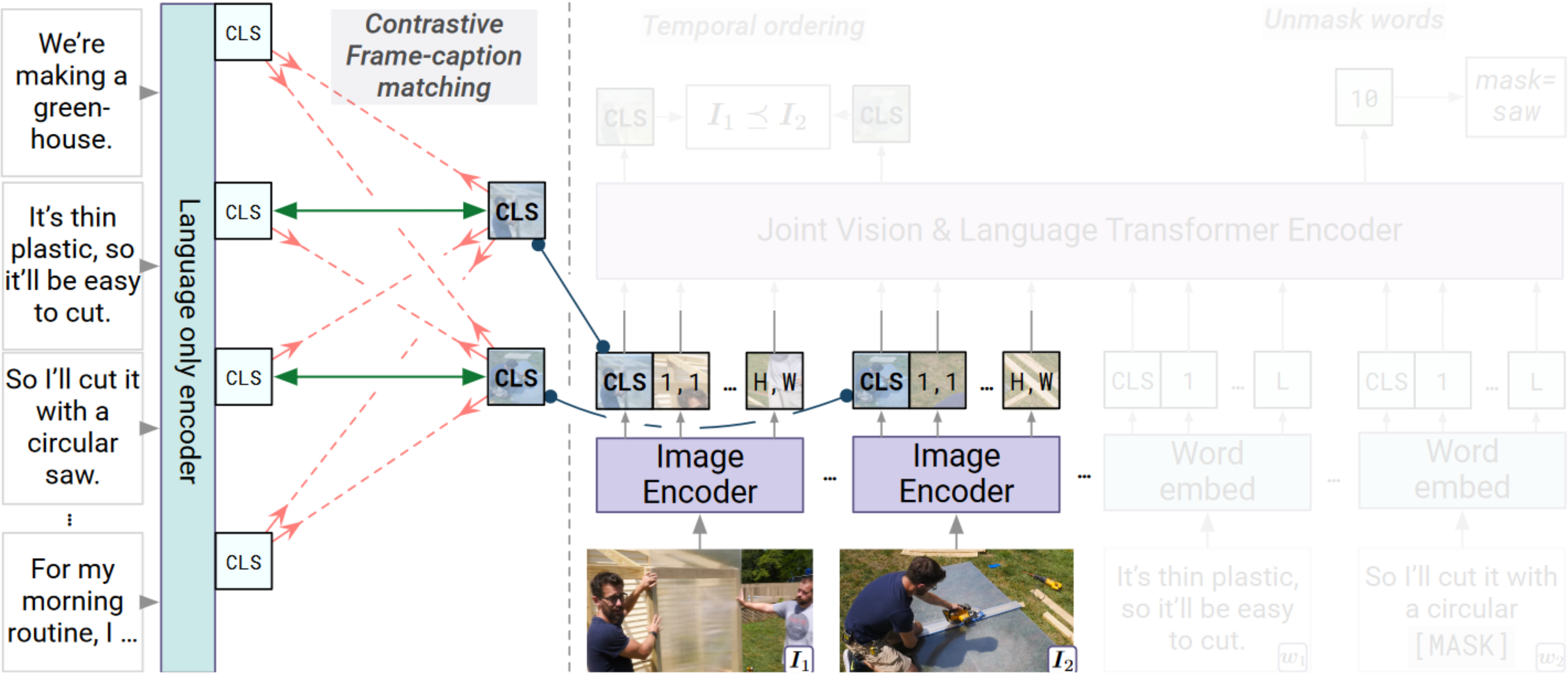
- **Key Idea: Learning temporal reasoning ability through massive video data.**



Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT: Multimodal Neural Script Knowledge Models

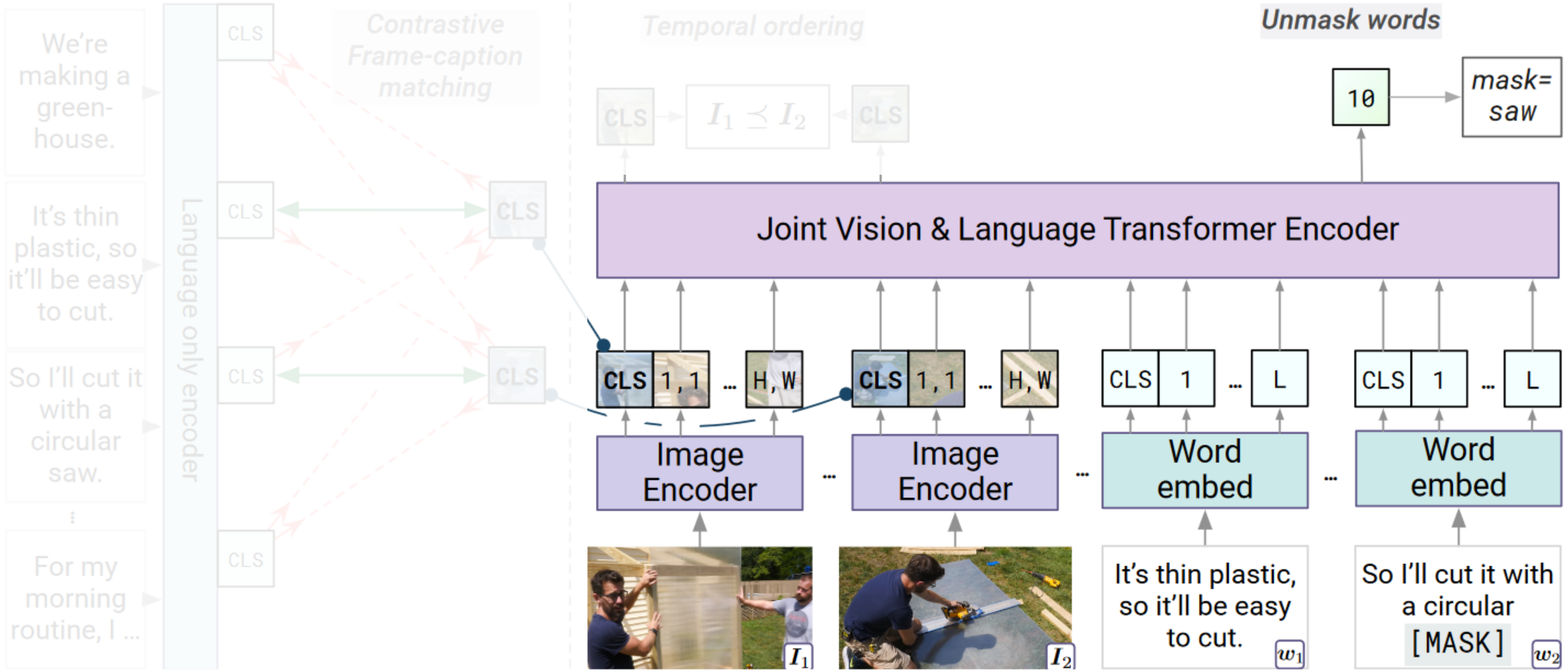
- Objective 1: Match between frame representations and text representations



Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT:
Multimodal Neural Script Knowledge Models

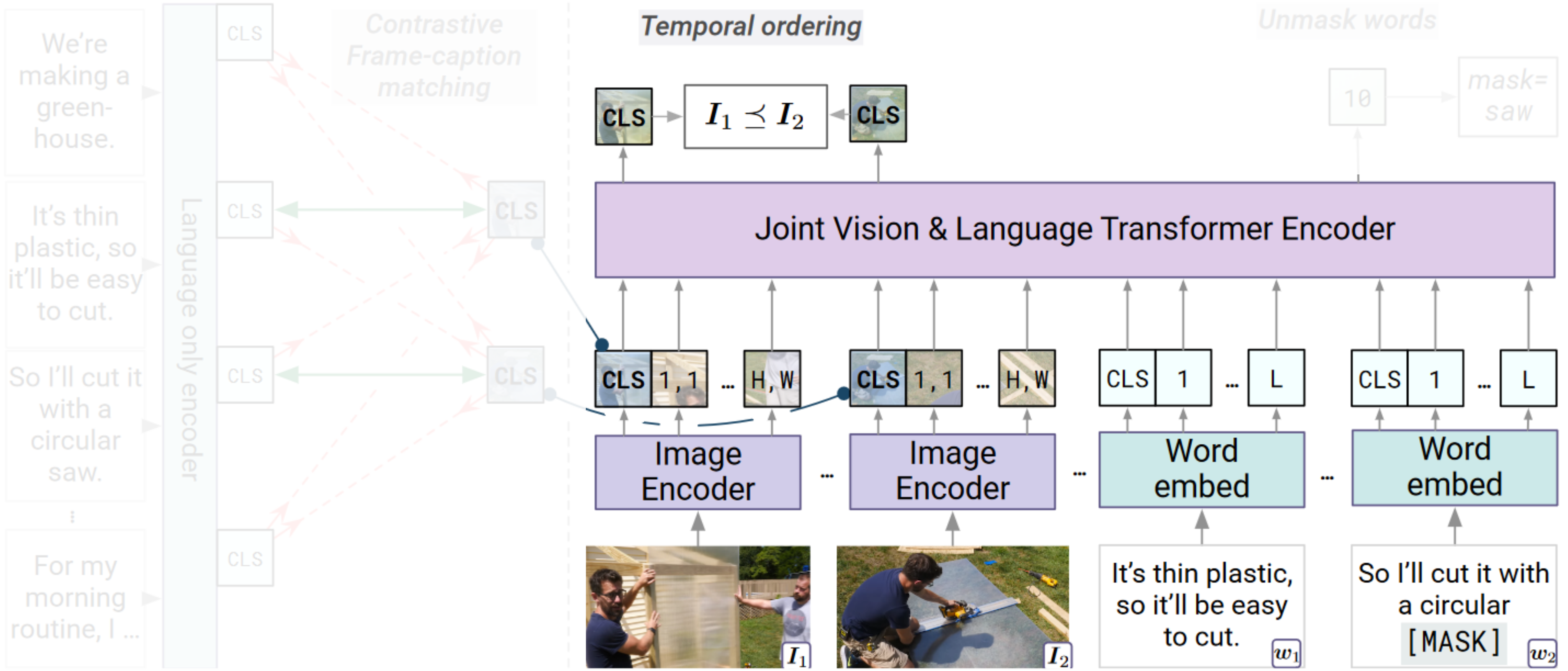
- Objective 2: Masked Token Modeling.



Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT:
Multimodal Neural Script Knowledge Models

- Objective 3: Temporal Ordering (Binary classification between each pair of frames).



Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT: Multimodal Neural Script Knowledge Models

- The model learns strong temporal reasoning ability and joint video-language reasoning ability.

Ordering Images from Visual Stories

	Spearman (↑)	Pairwise acc (↑)	Distance (↓)
CLIP [89]	.609	78.7	.638
UNITER [22]	.545	75.2	.745
MERLOT	**.733**	**84.5**	**.498**

State-of-the-art over various video-language tasks

Tasks	Split	Vid. Length	ActBERT [127]	ClipBERT$_{8x2}$ [67]	SOTA	MERLOT
MSRVTT-QA	Test	Short	-	37.4	41.5 [118]	**43.1**
MSR-VTT-MC	Test	Short	88.2	-	88.2 [127]	**90.9**
TGIF-Action	Test	Short	-	82.8	82.8 [67]	**94.0**
TGIF-Transition	Test	Short	-	87.8	87.8 [67]	**96.2**
TGIF-Frame QA	Test	Short	-	60.3	60.3 [67]	**69.5**
LSMDC-FiB QA	Test	Short	48.6	-	48.6 [127]	**52.9**
LSMDC-MC	Test	Short	-	-	73.5 [121]	**81.7**
ActivityNetQA	Test	Long	-	-	38.9 [118]	**41.4**
Drama-QA	Val	Long	-	-	81.0 [56]	**81.4**
TVQA	Test	Long	-	-	76.2 [56]	**78.7**
TVQA+	Test	Long	-	-	76.2 [56]	**80.9**
VLEP	Test	Long	-	-	67.5 [66]	**68.4**

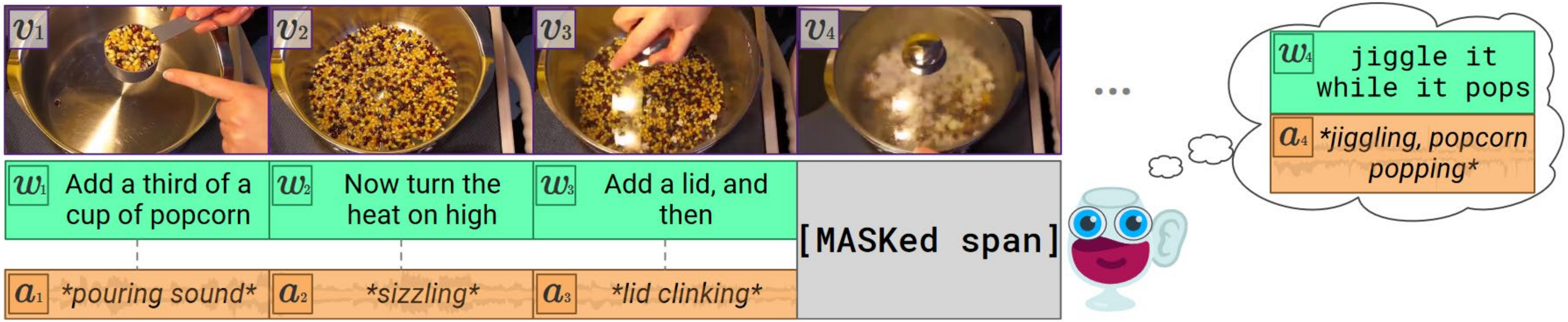
Predict future event given historical videos

- Limitation: short temporal span; importance of the temporal ordering loss is unclear.

Zellers, Rowan, et al. "Merlot: Multimodal neural script knowledge models." *Advances in Neural Information Processing Systems* 34 (2021): 23634-23651.

MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound

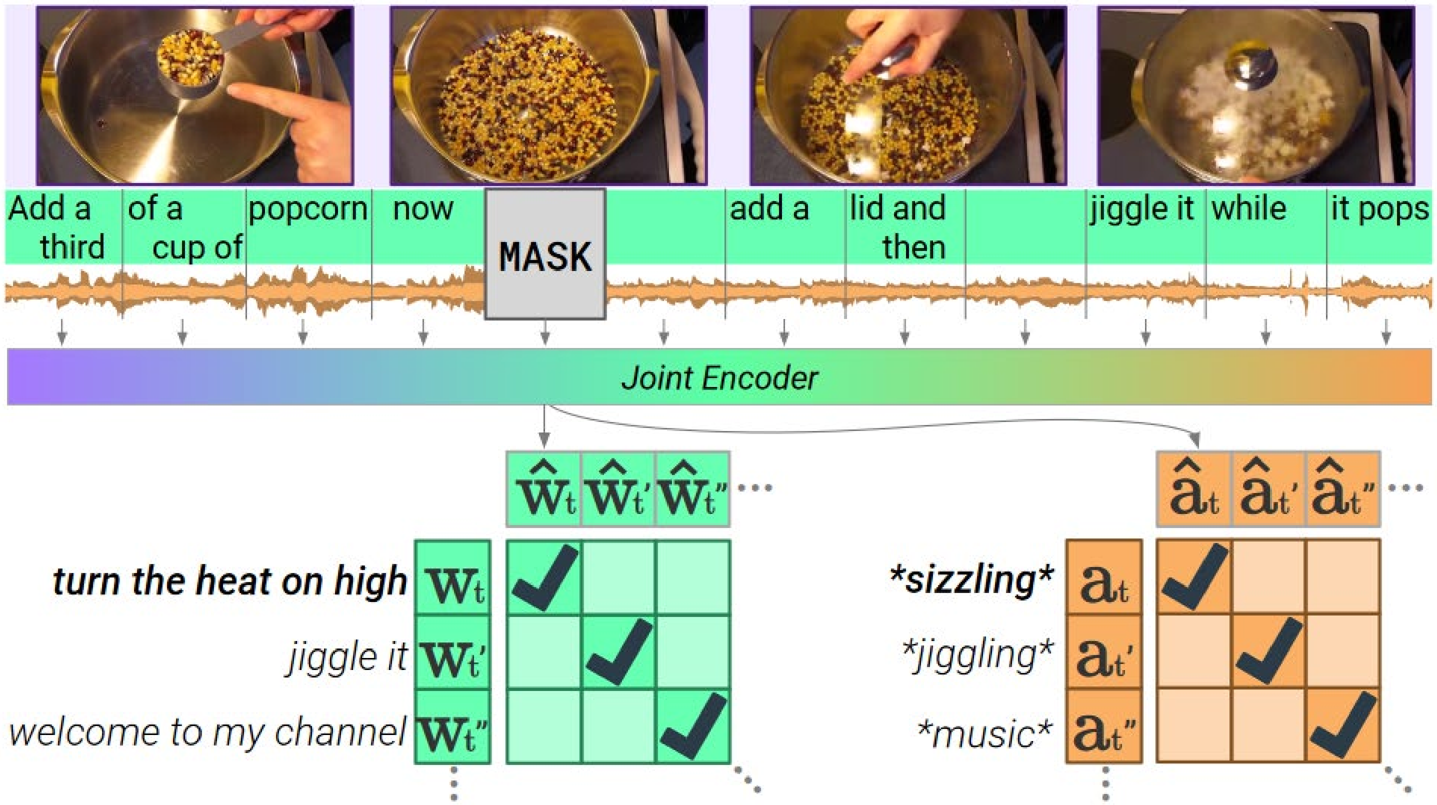
- Key Idea: Jointly learn script knowledge with video, language and audio.



Zellers, Rowan, et al. "Merlot reserve: Neural script knowledge through vision and language and sound." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound

- Key objective design: <u>contrastive loss</u> between predicted and actual representation of the masked audio/text



Zellers, Rowan, et al. "Merlot reserve: Neural script knowledge through vision and language and sound." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound

- Audio brings extra supervision and information towards stronger video understanding and video-language performance.
- Limitation: improvement on learned procedural knowledge may be less significant.

Action Recognition

	Model	Kinetics-600 (%) Top-1	Top-5
Vision Only	VATT-Base[2]	80.5	95.5
	VATT-Large [2]	83.6	96.6
	TimeSFormer-L [9]	82.2	95.6
	Florence [125]	87.8	97.8
	MTV-Base [122]	83.6	96.1
	MTV-Large [122]	85.4	96.7
	MTV-Huge [122]	89.6	98.3
	RESERVE-B	88.1	95.8
	RESERVE-L	89.4	96.3
+Audio	RESERVE-B	89.7	96.6
	RESERVE-L	**91.1**	**97.1**

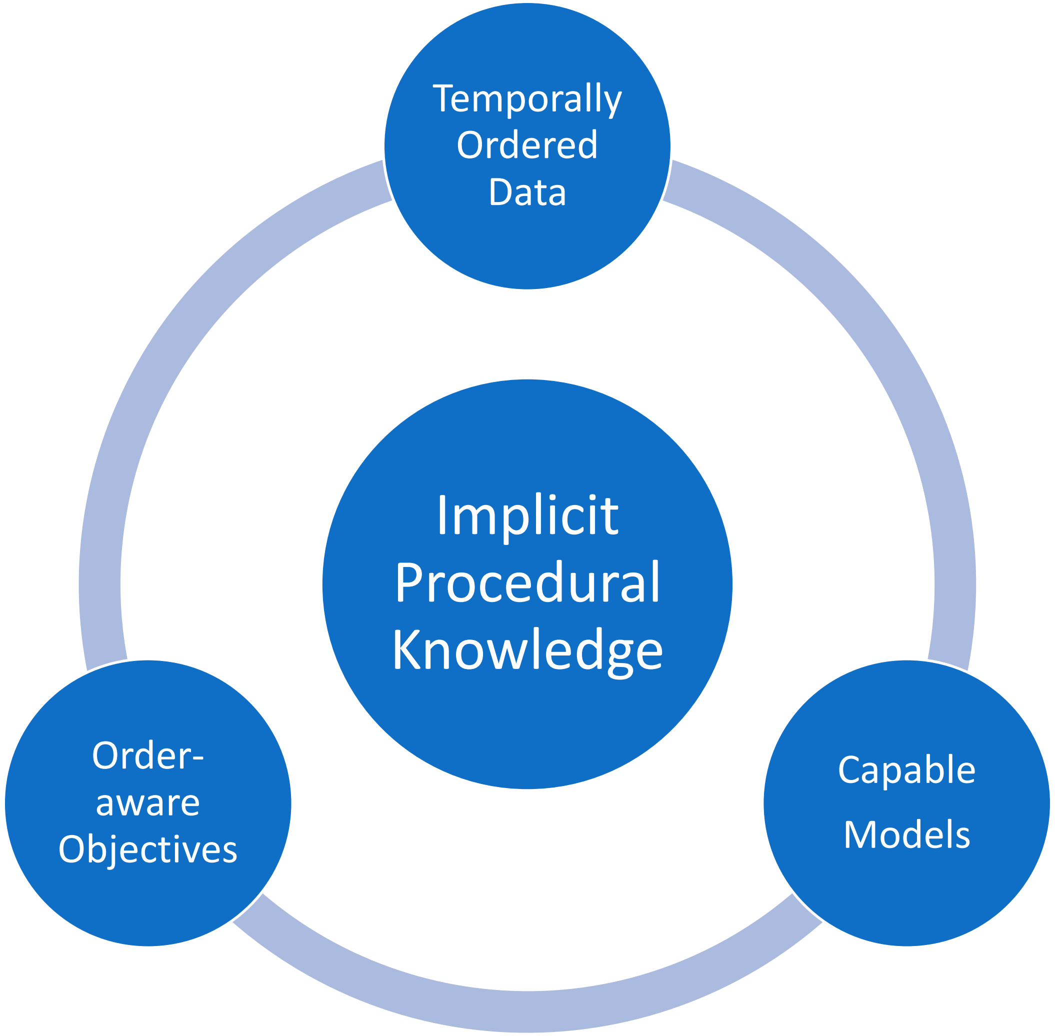
Requiring understand procedures of actions/objects

Situated Reasoning (STAR) (test acc; %)

	Model	Interaction	Sequence	Prediction	Feasibility	Overall
	Supervised SoTA (ClipBERT [74])	39.8	43.6	32.3	31.4	36.7
zero-shot	Random	25.0	25.0	25.0	25.0	25.0
	CLIP (VIT-B/16) [92]	39.8	40.5	35.5	36.0	38.0
	CLIP (RN50x16) [92]	39.9	41.7	36.5	**37.0**	38.7
	Just Ask (ZS)[123]					
	RESERVE-B	44.4	40.1	38.1	35.0	39.4
	RESERVE-L	42.6	41.1	37.4	32.2	38.3
	RESERVE-B (+audio)	**44.8**	**42.4**	**38.8**	36.2	**40.5**
	RESERVE-L (+audio)	43.9	42.6	37.6	33.6	39.4

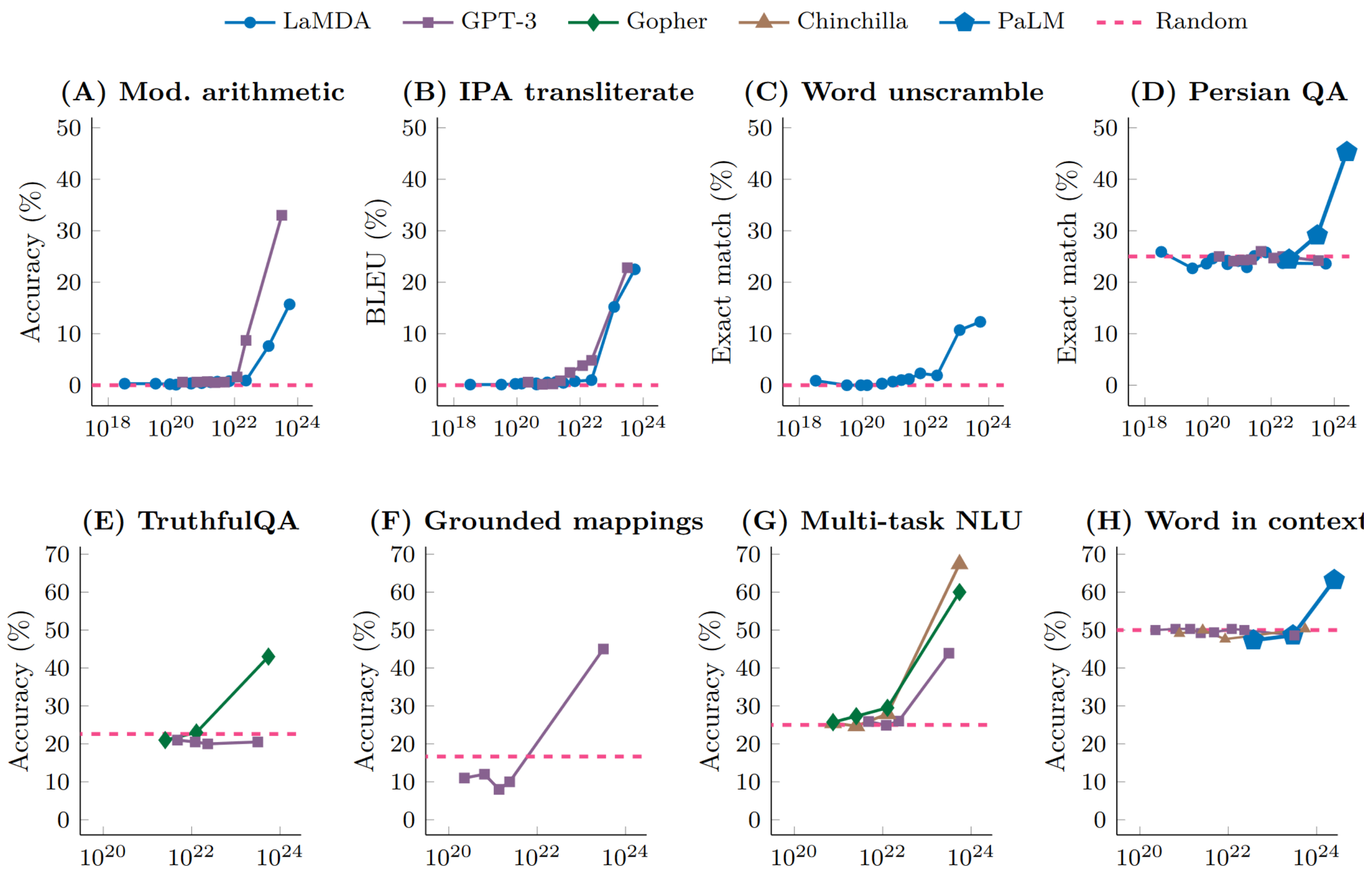
Zellers, Rowan, et al. "Merlot reserve: Neural script knowledge through vision and language and sound." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Summary of Methods Learning Implicit Knowledge



Future Challenge: Is there a critical point on scale?

- Can models learn procedural knowledge with a limited scale?

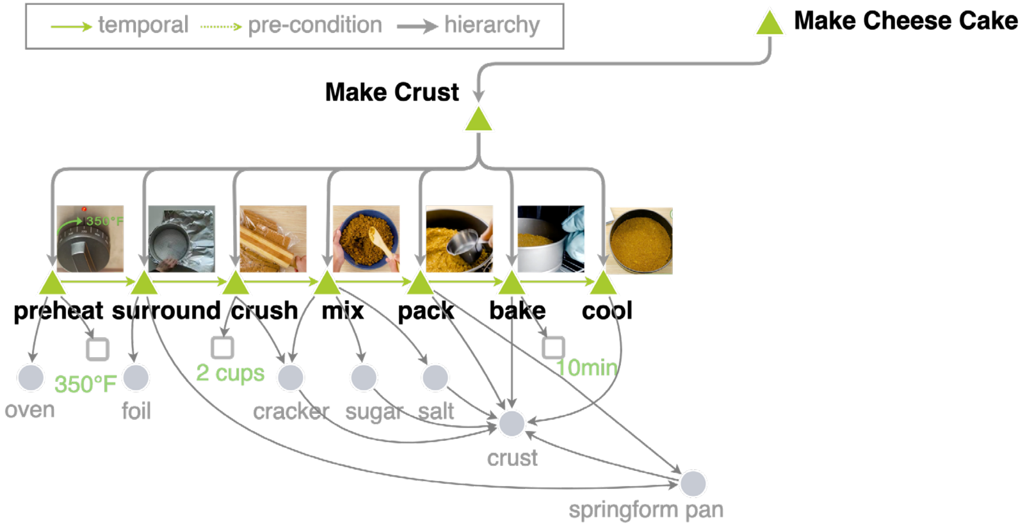


Many reasoning ability of <u>large language models</u> emerge when the model scale is larger than <u>a critical point</u>.

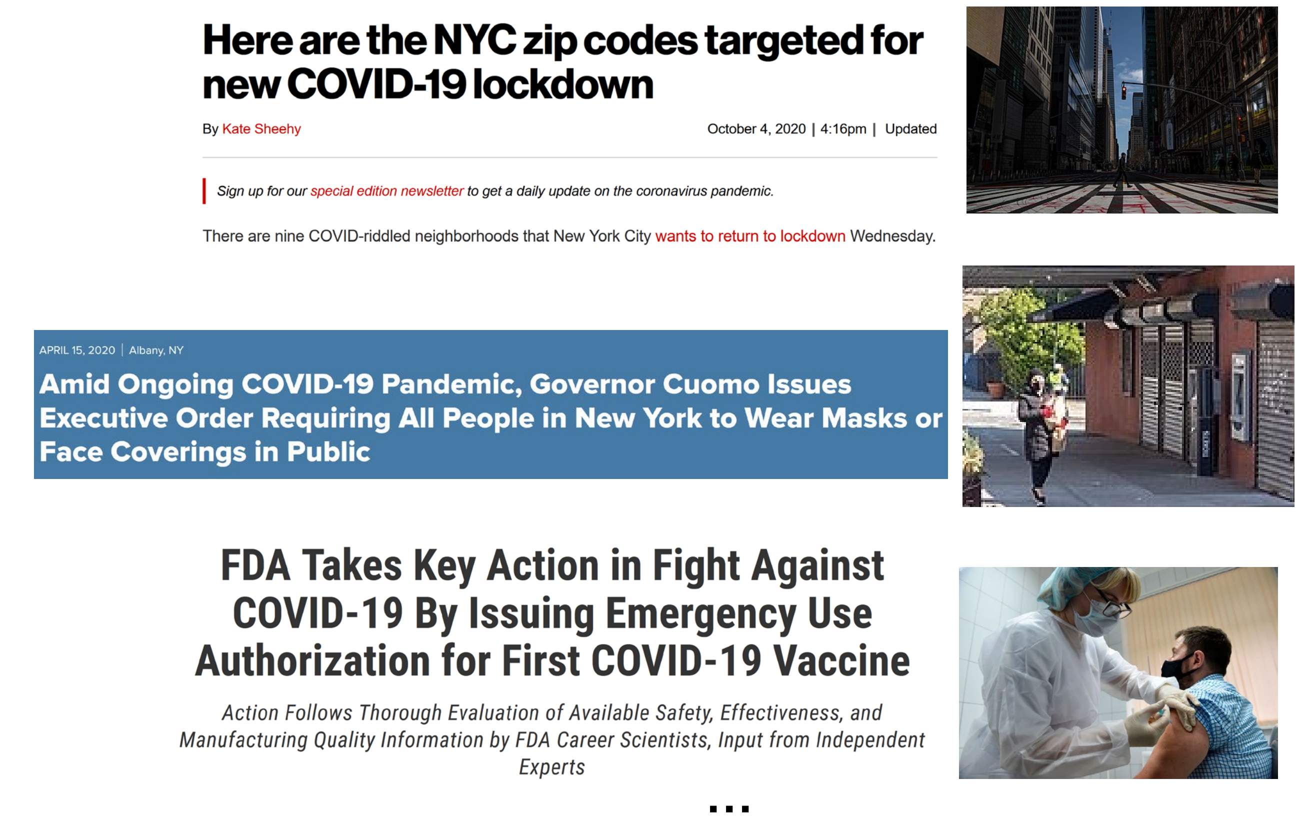
Wei, Jason, et al. "Emergent abilities of large language models." *arXiv preprint arXiv:2206.07682* (2022).

Future Challenge: From an instance to a set

- Can models learn from temporally ordered sets of instances?



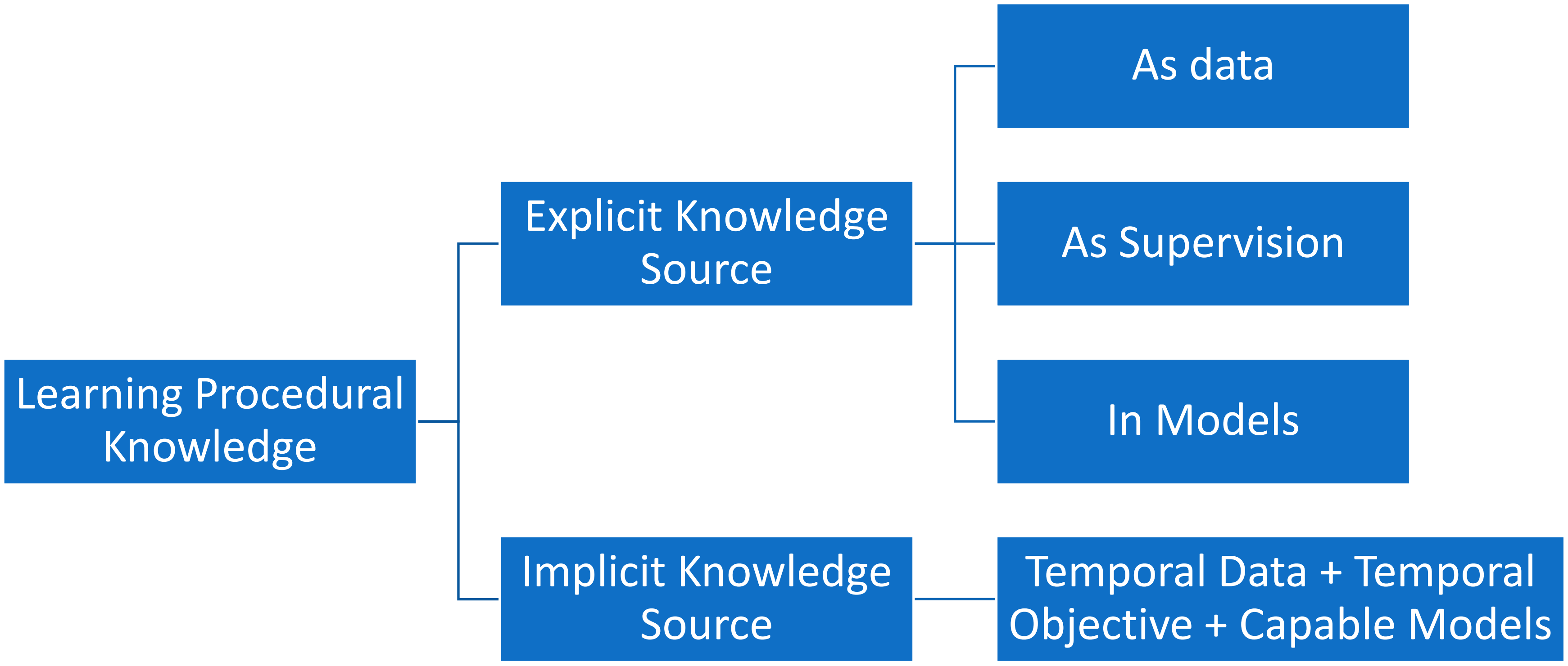
Current Data Instance



Real-world complex task

https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_City#cite_note-48
https://nypost.com/2020/10/04/here-are-the-nyc-zip-codes-targeted-for-new-covid-19-lockdown/
https://www.governor.ny.gov/news/amid-ongoing-covid-19-pandemic-governor-cuomo-issues-executive-order-requiring-all-people-new
https://www.fda.gov/news-events/press-announcements/fda-takes-key-action-fight-against-covid-19-issuing-emergency-use-authorization-first-covid-19
https://www.nature.com/articles/d41586-020-02684-9

Take-away Messages



Future Challenges:

From a sequence to a graph;

Interpret but not memorize;

Learning with limited scale;

From an instance to a set;

…

Q & A