Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

Carina Silberer

IMS, University of Stuttgart

CL Fall School 2024, Passau
16 - 20 September 2024

Outline

Outline

Action–Effects

Reasoning over Actions

- Actions can manipulate objects and change their states
- Reasoning about actions and their effects is helpful for action planning, explanation, prediction

Action–Effects

Reasoning over Actions

- Actions can manipulate objects and change their states
- Reasoning about actions and their effects is helpful for action planning, explanation, prediction

What/Where is the Contribution of Visual Information?

Reasoning Type (with corresponding Inputs and Outputs)	Example
Temporal Prediction	
Initial State, Action(s) → Resulting State	Potato peeled and cut into pieces, Frying → Potato pieces are crisp and golden
Temporal Explanation	
Initial State*, Resulting State → Explanation about Action(s)	Carrot with the skin removed → Peeling action performed on the whole carrot causes skin removal
Action(s), Resulting State → Explanation about Initial State	Freezing, Ice → Water if freezed turns into ice
Goal driven Action Prediction/Planning	
Initial State*, Resulting State → Action(s)	Cake batter in the baking pan, The cake is baked → Put the pan into oven for 20-25 minutes at 350F
Temporal Dependency	
Action(s) → Action(s) performed Before/After	Changing of a tire → Flattening of the tire (after) Eating a banana → Peeling the banana (before)

[Sampat et al., 2022]

Action–Effects

Reasoning over Actions

- Actions can manipulate objects and change their states
- Humans gain a vast amount of knowledge about actions

What/Where is the Contribution of Visual Information?

[Sampat et al., 2022]

Action–Effects

Reasoning over Actions

- Actions can manipulate objects and change their states
- Humans gain a vast amount of knowledge about actions

What/Where is the Contribution of Visual Information?

Knowledge Type	Example
Preconditions	**Liquid** objects can undergo *pouring*
Co-occurrence	**Knife** is often mentioned with *cut*
Locations	*Brushing* takes place in **washroom**
Tools/Materials	**Oven** is useful for *baking*
Object attributes	*Painting* an object changes its **color**
Intents	One *drinks* water if they are **thirsty**

[Sampat et al., 2022]

Action–Effect Modelling

Reasoning over Actions: Effects and Conditions

- Infinite number of combinations of objects, actions and effects
- Fine-grained visual differences in effects
- High language variation in actions and effects (states)
- Work in computer vision: action–effect prediction, state discovery, action–state composition learning

$\Rightarrow$ High-quality, large-scale training data

Action–Effects: Action Reasoning in V+L

[Isola et al., 2015]

Discovering States and Transformations in Image Collections

- **Goal:** Learn object states and transformations that can generalise over different nouns
  e.g., *melted* chocolate $\Leftrightarrow$ *melted* butter
- Test zero-shot / transfer learning: generalise to adjective+unseen noun compositions

Dataset

- Vocabulary of (adjective, antonym, noun) triplets,
  e.g., *(fresh, moldy, tomato)*
- Images collected by web querying with adj+noun pairs
- 115 adjectives denoting physical transformations
- 249 nouns denoting physical objects

Action–Effects: Action Reasoning in V+L

[Isola et al., 2015]

Discovering States and Transformations in Images: Tasks

Inference on unseen nouns, states are represented by adjectives.

1. **Image state classification**
   - Given an image I, predict the state(s) A of the shown (unseen) object
   - Trained with logistic regression for each adjective A: $g(A|I) = \sigma(-w_A^T f(I))$

w_A weight vector trained with LR, $f(I)$ visual CNN feature vector

2. **Collection parsing**

Action–Effects: Action Reasoning in V+L

[Isola et al., 2015]

Discovering States and Transformations in Images: Tasks

Inference on unseen nouns, states are represented by adjectives.

1. **Image state classification**
   - Given an image I, predict the state(s) A of the shown (unseen) object
2. **Collection parsing**
   - Given a set of images, assign a state to each image
   - Approach: Conditional Random Field, leverages state similarity between image pairs
   - Results: mean accuracy = 12%

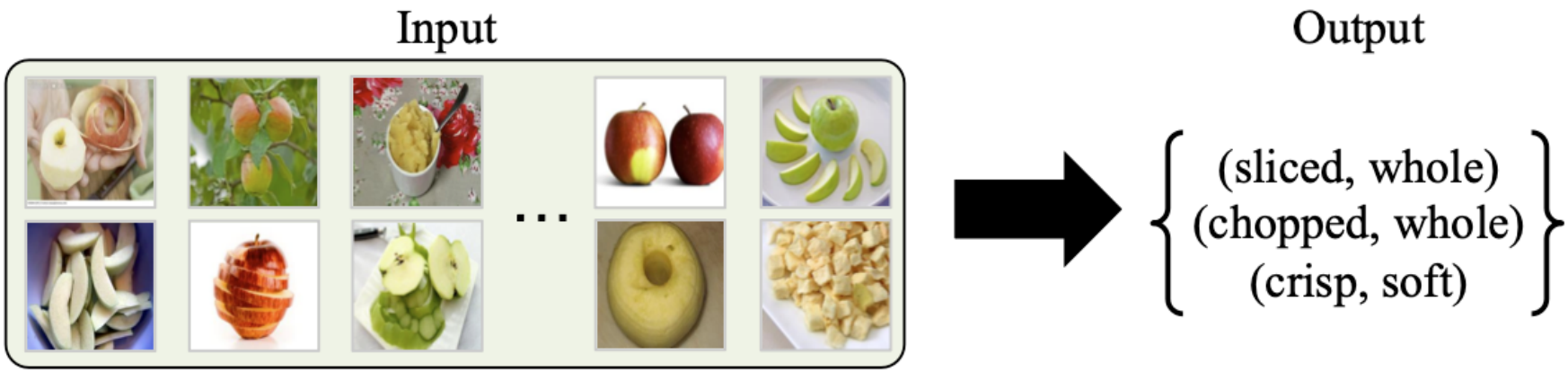
Action–Effects: Action Reasoning in V+L

[Isola et al., 2015]

Discovering States and Transformations in Images: Tasks

Inference on unseen nouns, states are represented by adjectives.

3. **Transformation prediction**:
   - Given a set of images of an object (noun), which state transformations (adjective set $\{\mathcal{A}_J\}$) are shown?
   - $$J = \arg\max_{J'} \sum_{j \in J'} \sum_{i \in \mathcal{I}} [e^{\lambda g(\mathcal{A}_J | I_i)} + e^{\lambda g(ant(\mathcal{A}_j) | I_i)}]$$
   - Results: mAP = 0.39



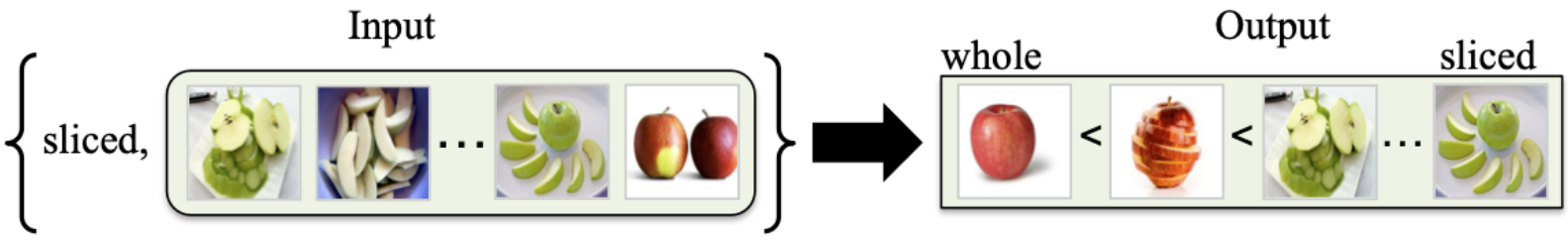
Action–Effects: Action Reasoning in V+L

[Isola et al., 2015]

Discovering States and Transformations in Images: Tasks

Inference on unseen nouns, states are represented by adjectives.

4. **Transformation ordering**
   - Given a set of images and state A, order it according to the transformation from state to opposite state:
   - $g(ant(A)|I_i) < \cdots < g(A|I_i)$
   - results: mean $\rho = 0.46$



Action–Effects: Action Reasoning in V+L

[Isola et al., 2015]

Discovering States and Transformations in Images: Review

- Approach based on words (symbols), synonyms were not taken into account which weakens the knowledge transfer / zero-shot claim
- Q: Does polysemy/ambiguity of adjectives play a role in the ability to learn to transfer state knowledge?

Action–Effects: Action Reasoning in V+L

[Tateno et al., 2024]

Learning Object States from Actions via Large Language Models

- How does the state of objects change as a result of an action?
- How can object state (changes) be learnt from videos? (!reporting bias)

$\Rightarrow$ Infer object states by LLMs from the actions described in given video narrations

- Assumption:
  - Action–object state relationships are implicitly encoded in LLMs
  - State can be inferred from action and the object's past state

Action–Effects: Action Reasoning in V+L

[Tateno et al., 2024]

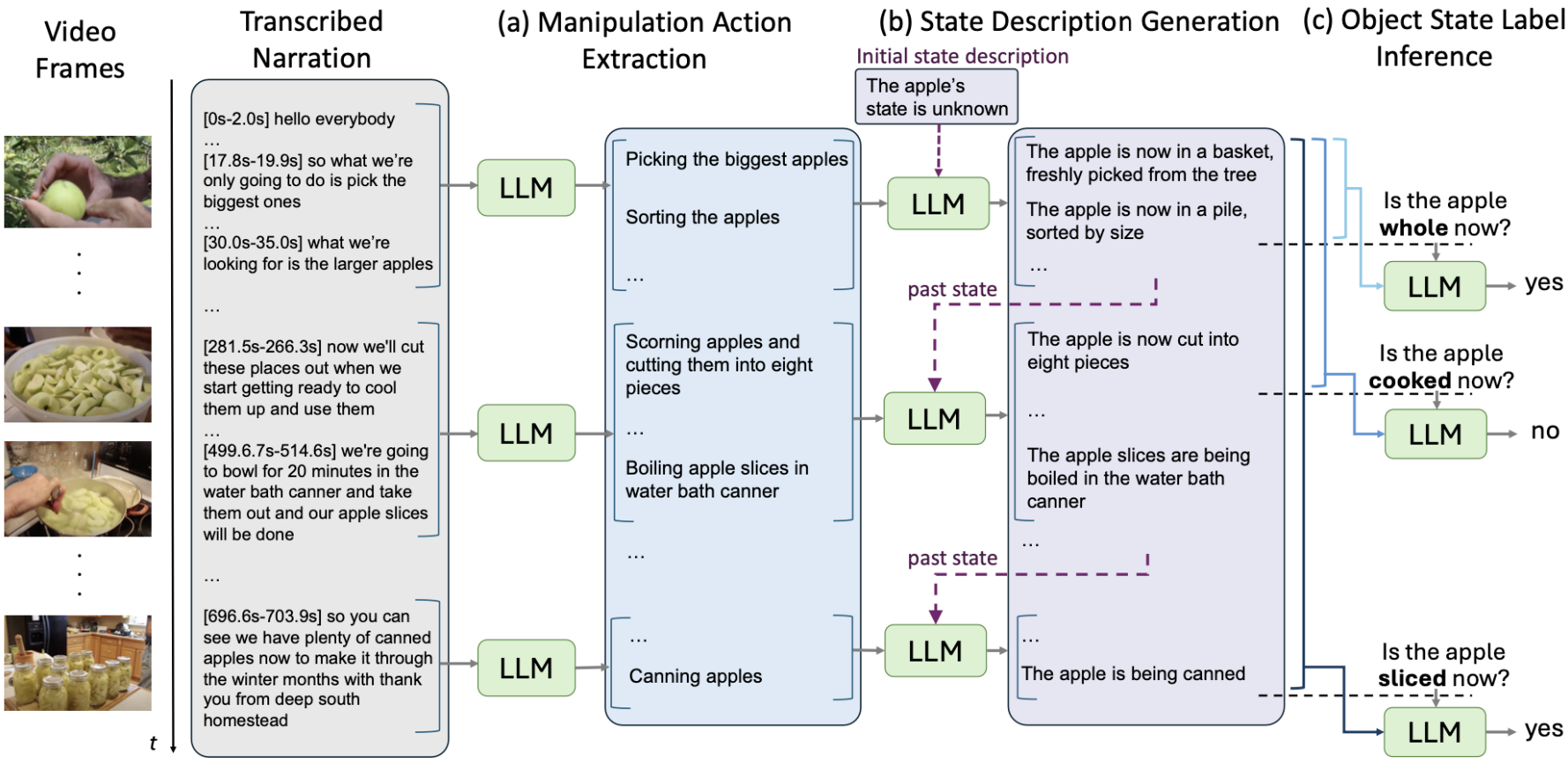
Learning Object States from Actions via LLMs

- **Goal:** Given video with narrations, predict the frames with a specific object state
- **Task:** Multi-label frame-wise classification
- **Approach**
  1. Manipulation action extraction
  2. State description generation from actions
  3. Context-aware object state label inference

Action–Effects: Action Reasoning in V+L

[Tateno et al., 2024]

Approach



Action–Effects: Action Reasoning in V+L

[Tateno et al., 2024]

Step 1: Manipulation action extraction

- Use LLM to extract manipulation actions from video narration

> Analyze a segment of video transcript provided in CSV format. The CSV only has one column and no headers.
> {block of transcribed narration}
>
> You need to list and describe all object manipulating actions performed in the video in detail. Do not include actions such as greeting, thanking, explaining or summarizing that do not manipulate any object. Do not summarize actions too short, but make sure you describe all the actions in each sentence in detail. Especially, make sure to use original nouns (object names) and verbs (human actions) when you summarize.
>
> In addition, for each action, extract the part of the transcript that describes or supports the action. Make sure to extract the whole sentence for support. When you need to combine multiple lines from the transcript to support an action, separate them with a space instead of a comma or line break.
>
> The answer format should be in CSV format. Make sure to use quotation marks for each action and the part of the transcript.
> Format: "<detailed summary of action>","<part of the transcript (This should be exactly the same as the original. Don't skip.)>"
> Example: "Adding whisked eggs into the pan.","let's add the whisked eggs into the pan"

Action–Effects: Action Reasoning in V+L

[Tateno et al., 2024]

Step 2: Object state description generation

- Given sequence of actions and object name, generate sequentially a sentence describing the corresponding state (resulting from the action) *and* the object name

$\Rightarrow$ Object state tracking and name tracking

"Stirring the mixture" $\rightarrow$ *The almond flour and baking soda are mixed in a bowl*

> You will be given a sequence of actions. Trace the history of changes in the internal state of {object} and describe it in detail for each action.
> The initial state of the {object} is "{previous state description}". You don't need to include the initial state in the answer.
>
> The answer format should be in CSV with the action column and state description column. Make sure that each state description includes the whole history of what has been done on the {object} so far. The description should be a complete sentence starting with "The {object}", but do not finish only with this.
> If the internal state doesn't change after the action, you don't have to change the state description from the previous one. Use quotation marks for the description. The answer format:

Action–Effects: Action Reasoning in V+L

[Tateno et al., 2024]

Step 3: Context-aware object state label inference

- Prompt LLM for presence of each given object state (*yes*, *no*, *ambiguous*)
- Prompt includes all previous+current state description + state definitions

"Is the apple sliced now?" $\rightarrow$ *yes*

"The apple is sliced" refers to a state ... been cut into thin or narrow pieces ...

This is a history of state of {object}:
{list of state description up to that point}

Now, does the state of {object} fit the definition of "{state}"?

Object state definition:
{definition}

Think step-by-step as follows:
- First, list all points for judging the state "{state}" from the object state definition. Make sure to describe in detail.
- Second, carefully compare all listed judging points to the whole history of the object state by tracing it in detail.
- Then, answer Yes/No about whether the current state of {object} is

Action–Effects: Action Reasoning in V+L

[Tateno et al., 2024]

Step 3: Context-aware object state label inference

This is a history of state of {object}:
{list of state description up to that point}

Now, does the state of {object} fit the definition of "{state}"?

Object state definition:
{definition}

Think step-by-step as follows:
- First, list all points for judging the state "{state}" from the object state definition. Make sure to describe in detail.
- Second, carefully compare all listed judging points to the whole history of the object state by tracing it in detail.
- Then, answer Yes/No about whether the current state of {object} is consistent with the definition and give detailed reasons. If the history doesn't contain enough information for judging, answer Ambiguous.

Make sure to answer the three things above in detail, separating them by newline as follows:
Judging points: [judging points from object state definition]

Comparison: [comparison]

Answer: [yes/no/ambiguous and why]

Screenshot

Action–Effects: Action Reasoning in V+L

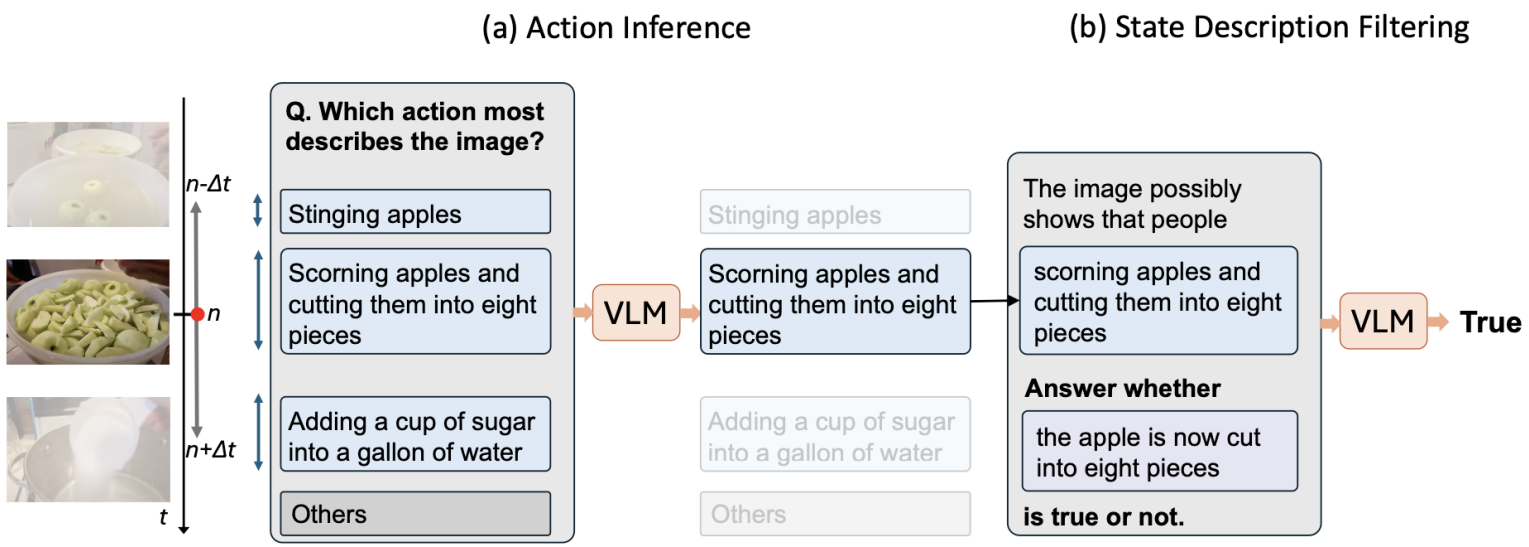
[Tateno et al., 2024]

Interval alignments with Vision–Language Models

- Potential temporal misalignments of actions and object states between extracted actions+states and video frames

(a) Given frame, which is the most plausible action among action candidates?

(b) Use state description to discard irrelevant frames, assign states to relevant frames

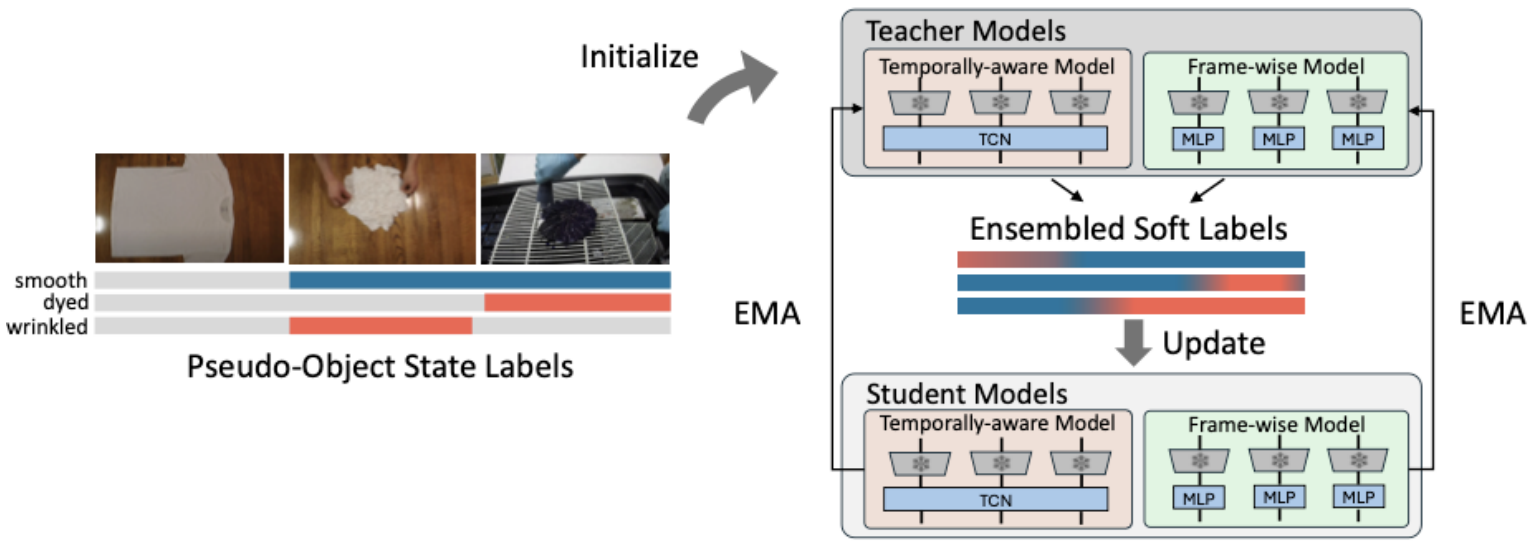


Action–Effects: Action Reasoning in V+L

[Tateno et al., 2024]

Learning from pseudo-object state labels

- Given frames, each labeled with the inferred action and object states
- Pseudo-label preprocessing: Assign negative labels to frames without target objects (using CLIP)
- Self-training of frame-based state prediction via ensemble teacher models:

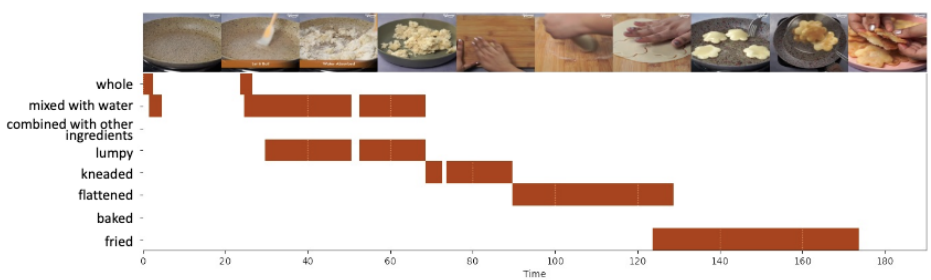


Action–Effects: Action Reasoning in V+L

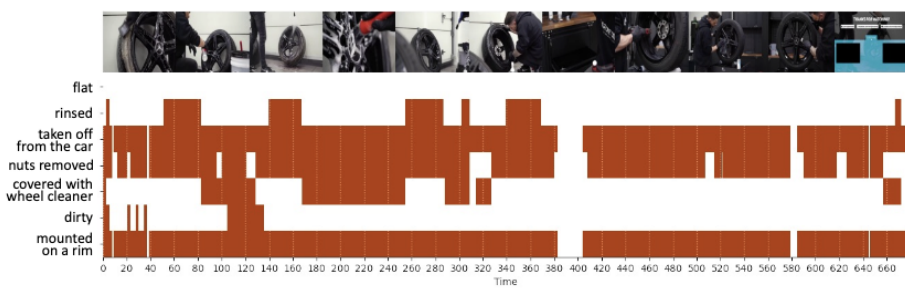
[Tateno et al., 2024]

Experiments

- Test data: MOST
  - 61 manually labeled YouTube videos
  - six object categories, 10 states each
- Training data: Howto100M
- LLM: GPT3.5, VLLM: LLaVa-1.5 13B



(a) Scene: Cooking food with flour.



(b) Scene: Washing car tire.

Action–Effects: Action Reasoning in V+L

[Tateno et al., 2024]

Results: MOST

Method	Apple		Egg		Flour		Shirt		Tire		Wire		**Average**	
	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP
LLaVA [25]	0.34	—	0.29	—	0.35	—	0.28	—	0.47	—	0.27	—	0.33	—
CLIP [34]	0.42	0.35	0.37	0.28	0.38	0.26	0.33	0.27	0.55	0.45	0.33	0.25	0.39	0.31
InternVideo [45]	0.46	0.39	0.44	0.39	0.43	0.36	0.40	0.32	0.57	0.45	0.40	0.31	0.45	0.37
Ours	**0.53**	**0.50**	**0.53**	**0.48**	**0.55**	**0.49**	**0.50**	**0.45**	**0.61**	**0.52**	**0.50**	**0.42**	**0.54**	**0.48**

Comparison zero-shot models:

Input: image (video frame) + object state definition

- LLaVA-1.5: prompted to describe frame, then to answer if state is present in image
- CLIP: compute similarity between image and state definition
- InternVideo: video-based, similar to CLIP

References I

- Implicit Affordance Acquisition via Causal Action–Effect Modeling in the Video Domain [Yang and Silberer, 2023]
- Beyond Seen Primitive Concepts and Attribute-Object Compositional Learning [Saini et al., 2024]
- Learning the Effects of Physical Actions in a Multi-modal Environment [Dagan et al., 2023]
- Commonsense Justification for Action Explanation [Yang et al., 2018]
- hat Action Causes This? Towards Naive Physical Action-Effect Prediction [Gao et al., 2018]
- Survey [Sampat et al., 2022]
- (NLP) Everything Happens for a Reason: Discovering the Purpose of Actions in Procedural Text [Dalvi et al., 2019]
- (NLP) PIQA: Reasoning about Physical Commonsense in Natural Language [Bisk et al., 2020]

References II

- (NLP) Extraction of Common-Sense Relations from Procedural Task Instructions using BERT [Losing et al., 2021]

References I

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., and Choi, Y. (2020).
Piqa: Reasoning about physical commonsense in natural language.
In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Dagan, G., Keller, F., and Lascarides, A. (2023).
Learning the effects of physical actions in a multi-modal environment.
In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 133–148.

Dalvi, B., Tandon, N., Bosselut, A., Yih, W.-t., and Clark, P. (2019).
Everything happens for a reason: Discovering the purpose of actions in procedural text.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4496–4505.

Gao, Q., Yang, S., Chai, J. Y., and Vanderwende, L. (2018).
What action causes this? towards naive physical action-effect prediction.
In *Annual Meeting of the Association for Computational Linguistics*.

Isola, P., Lim, J. J., and Adelson, E. H. (2015).
Discovering states and transformations in image collections.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391.

Losing, V., Fischer, L., and Deigmöller, J. (2021).
Extraction of common-sense relations from procedural task instructions using BERT.
In *Proceedings of the 11th Global Wordnet Conference*, pages 81–90.

Saini, N., Pham, K., and Shrivastava, A. (2024).
Beyond seen primitive concepts and attribute-object compositional learning.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14466–14476.

References II

Sampat, S. K., Patel, M., Das, S., Yang, Y., and Baral, C. (2022).
Reasoning about actions over visual and linguistic modalities: A survey.
CoRR, abs/2207.07568.

Tateno, M., Yagi, T., Furuta, R., and Sato, Y. (2024).
Learning object states from actions via large language models.
arXiv e-prints, pages arXiv–2405.

Yang, H.-Y. and Silberer, C. (2023).
Implicit affordance acquisition via causal action–effect modeling in the video domain.
In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 846–871.

Yang, S., Gao, Q., Sadiya, S., and Chai, J. (2018).
Commonsense justification for action explanation.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2627–2637.