CVPR
JUNE 17-21, 2024
SEATTLE, WA

Why Not Use Your Textbook? Knowledge-Enhanced Procedure Planning of Instructional Videos

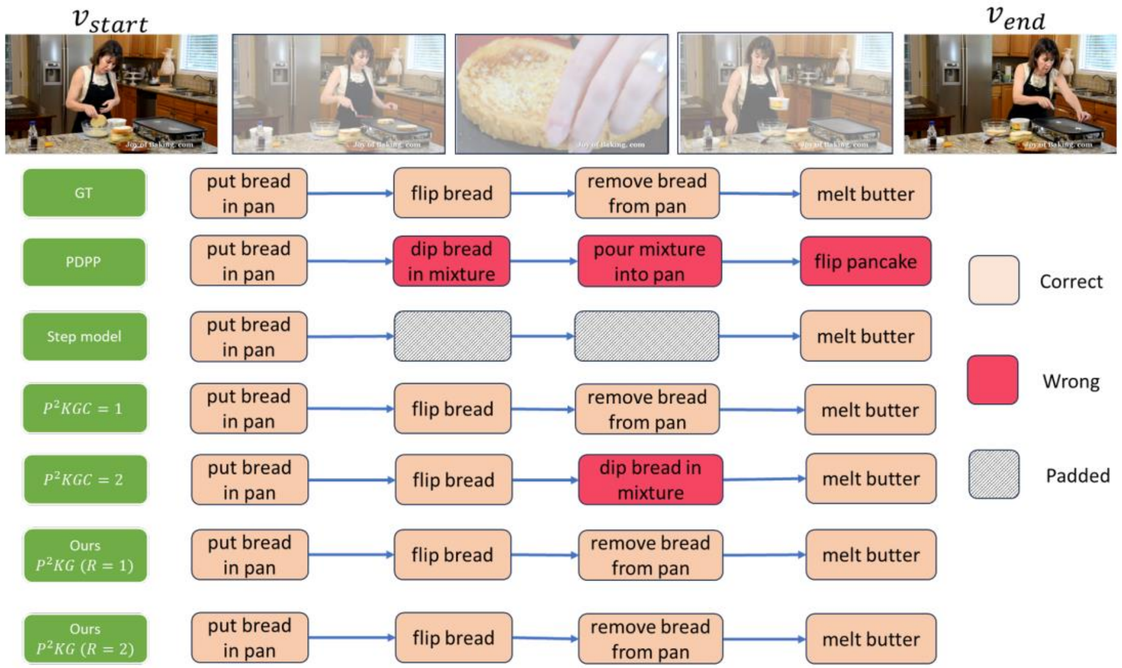
Kumaranage Ravindu Yasas Nagasinghe [1] , Honglu Zhou [2] , Malitha Gunawardhana [1,3]
Martin Renqiang Min [2] , Daniel Harari [4] , Muhammad Haris Khan [1]

[1]Mohamed bin Zayed University of Artificial Intelligence, [2]NEC Laboratories, USA,
[3]University of Auckland, [4]Weizmann Institute of Science

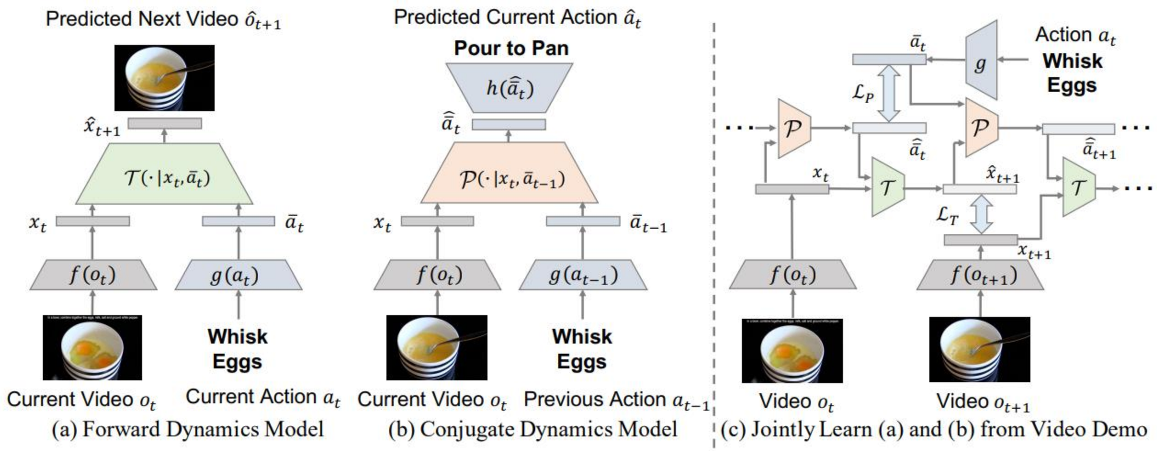
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE
NEC
NEC Laboratories America
UNIVERSITY OF AUCKLAND Waipapa Taumata Rau NEW ZEALAND
מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

What is Procedure Planning?

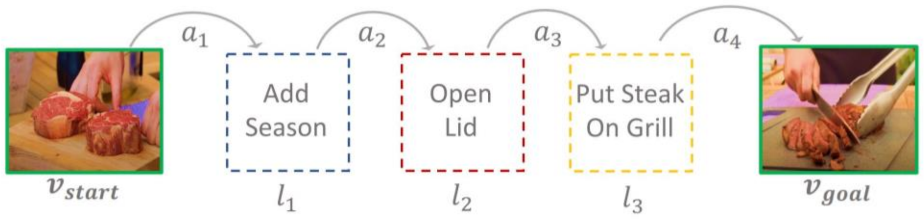
- Procedure planning in instructional videos requires an agent to create a sequence of actionable steps.
- This involves developing a plan that guides the transformation from an initial visual observation of the physical environment to reaching a desired goal state.



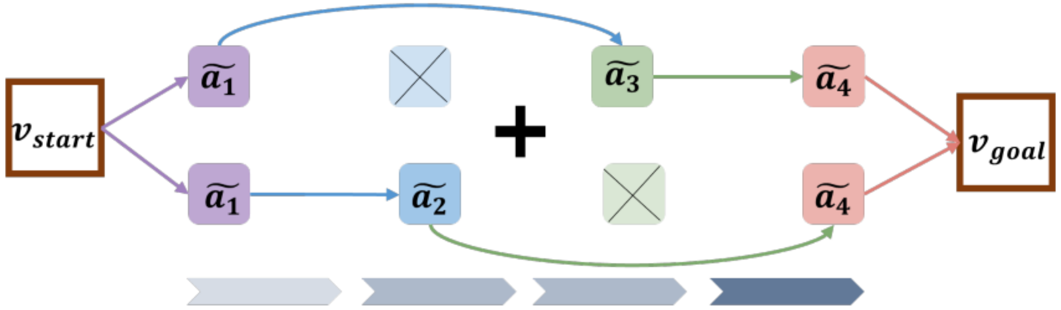
Prior work



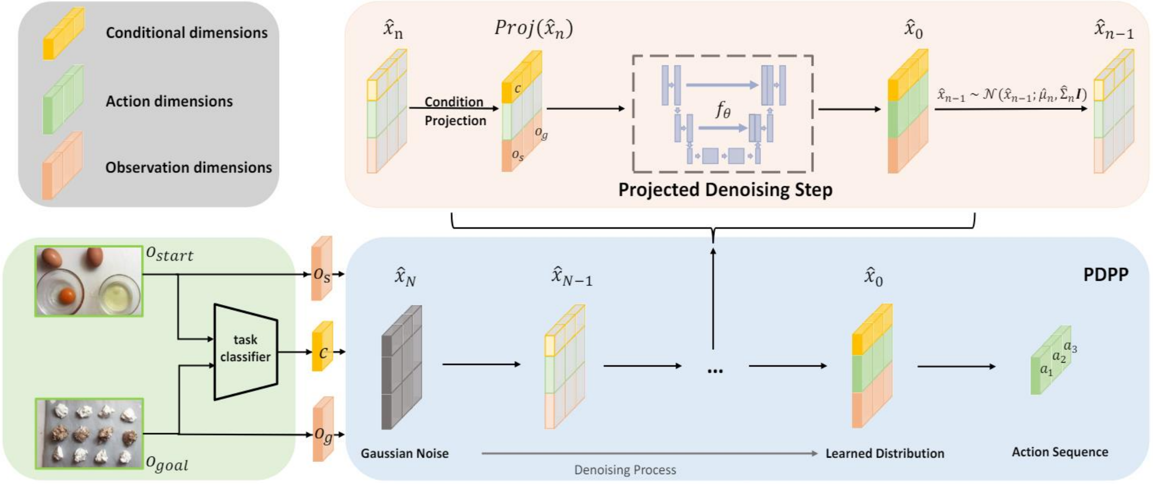
DDN



P3IV



Skip plan



PDPP

Motivation

- Procedure planning with minimal supervision while navigating the complexities of step sequencing and its potential variations.

- Challenges faced by prior work:
  - The presence of implicit temporal and causal constraints in the sequencing of steps.
  - The existence of numerous viable plans given an initial and a goal state.
  - Need to incorporate the real-life knowledge both in task-sharing steps and in managing the inherent variability in transition probabilities between steps.
  - Extensive use of annotations.

Contribution

- We propose KEPP, a Knowledge-Enhanced Procedure Planning system for instructional videos that leverages rich procedural knowledge from a probabilistic procedural knowledge graph (P^2KG).
- Requires only a minimal number of annotations for supervision.
- Decompose the problem in procedure planning of instructional videos into two sections.
- Experimental evaluations on three widely-used datasets.

Methodology

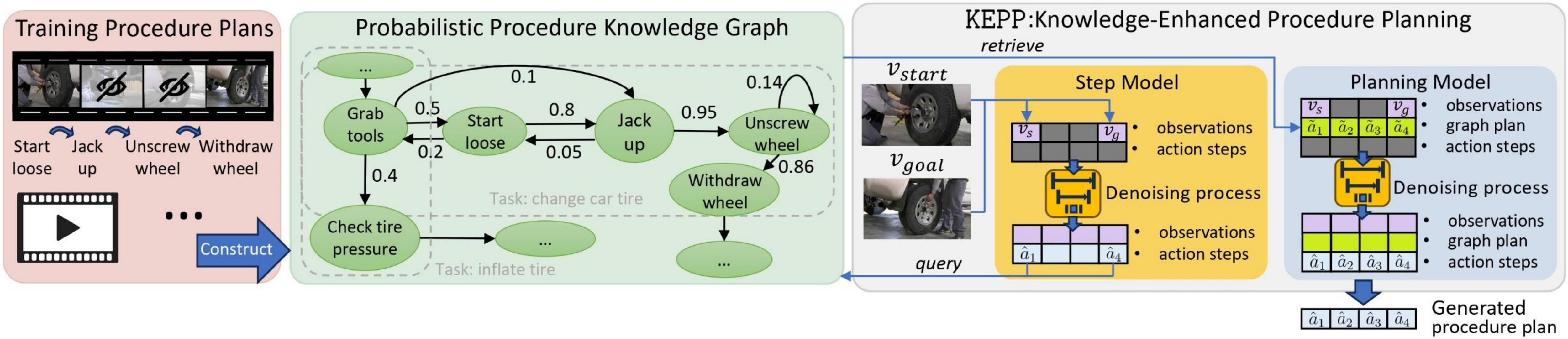
- Problem setup: $p(a_{1:T}|v_s, v_g)$
- Decompose the procedure planning problem:

$$\boxed{p(\hat{a}_{1:T}|v_s, v_g) = p(\hat{a}_{2:T-1}|\hat{a}_1, \hat{a}_T)p(\hat{a}_1, \hat{a}_T|v_s, v_g)}$$

- Harnessing a Probabilistic Procedural Knowledge Graph (P^2KG):

$$\boxed{p(\hat{a}_{1:T}|v_s, v_g) = p(\hat{a}_{1:T}|\tilde{a}_{1:T}, v_s, v_g)p(\tilde{a}_{1:T}|\hat{a}_1, \hat{a}_T)p(\hat{a}_1, \hat{a}_T|v_s, v_g)}$$

Methodology



- Identify the beginning and conclusion steps according to the input v_s and v_g using step model.
- Conditioned on these steps, query the graph to retrieve relevant P²KG plan $\tilde{a}_{1:T}$.
- Use the P²KG path conditioned planning model to generate procedure plan.

Methodology : Diffusion

- Step model:
  - Adapt a Conditioned Projected Diffusion Model to identify the first action step and the final step.

$$p_\theta\left(x_{n-1} | x_n\right) = \mathcal{N}\left(x_{n-1}; \mu_\theta\left(x_n, n\right), \Sigma_\theta\left(x_n, n\right)\right)$$

$$\begin{bmatrix} \hat{v}_1 & \hat{v}_2 & \ldots & \hat{v}_{T-1} & \hat{v}_T \\ \hat{a}_1 & \hat{a}_2 & \ldots & \hat{a}_{T-1} & \hat{a}_T \end{bmatrix} \xrightarrow{\text{Projection}} \begin{bmatrix} v_s & 0 & \ldots & 0 & v_g \\ \hat{a}_1 & 0 & \ldots & 0 & \hat{a}_T \end{bmatrix}$$

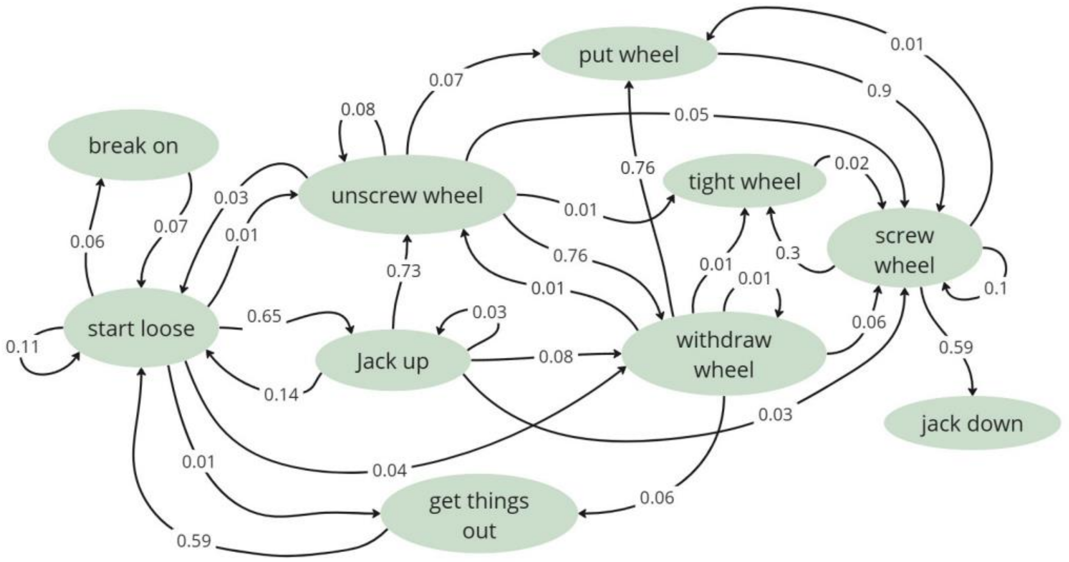
- Planning model:
  - Project operation keeps the dimensions of the visual state, P²KG recommendation, and zero-padding unaltered.

$$\begin{bmatrix} v_s & 0 & \ldots & 0 & v_g \\ \tilde{a}_1 & \tilde{a}_2 & \ldots & \tilde{a}_{T-1} & \tilde{a}_T \\ a_1 & a_2 & \ldots & a_{T-1} & a_T \end{bmatrix}$$

Methodology : Procedural knowledge graph

- P[2]KG = (V, E, w) is a directed and weighted graph created using the training data.
- Queries are made to the P[2]KG using the first ($\hat{a}_1$) and last ($\hat{a}_T$) actions predicted by the step model.
- Find the highest probable paths by multiplying the probability weights of the edges along the path.
- The top R paths are selected as the recommended procedure plans from the P[2]KG and are aggregated through linear weighting into a single path and given to plan model.

$$R = 1 \rightarrow weights : 1$$
$$R = 2 \rightarrow weights : \frac{2}{3}, \frac{1}{3}$$
$$R = 3 \rightarrow weights : \frac{3}{5}, \frac{1}{5}, \frac{1}{5}$$
$$R = 4 \rightarrow weights : \frac{4}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}$$
$$R = 5 \rightarrow weights : \frac{5}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}$$



Results : CrossTask

- We compare our model performance with prior works on the datasets CrossTask, COIN, and NIV.

Models	Required Annotations				$T=3$			$T=4$		
	step class	visual states	step text	task class	$SR^{\uparrow}$	$mAcc^{\uparrow}$	$mIoU^{\uparrow}$	$SR^{\uparrow}$	$mAcc^{\uparrow}$	$mIoU^{\uparrow}$
Random	✓				< 0.01	0.94	1.66	< 0.01	0.83	1.66
Retrieval-Based	✓				8.05	23.3	32.06	3.95	22.22	36.97
WLTDO [16]	✓	✓			1.87	21.64	31.70	0.77	17.92	26.43
UAAA [1]	✓	✓			2.15	20.21	30.87	0.98	19.86	27.09
UPN [45]	✓	✓			2.89	24.39	31.56	1.19	21.59	27.85
DDN [9]	✓	✓			12.18	31.29	47.48	5.97	27.10	48.46
PlaTe [46]	✓	✓			16.00	36.17	65.91	14.00	35.29	55.36
Ext-GAIL wo Aug. [7]	✓	✓			18.01	43.86	57.16	-	-	-
Ext-GAIL [7]	✓	✓			21.27	49.46	61.70	16.41	43.05	60.93
P^3IV ♣ [55]	✓		✓		23.34	49.96	73.89	13.40	44.16	70.01
PDPP ♣ [50]	✓			✓	26.38	55.62	59.34	18.69	52.44	62.38
E3P ♣ [49]	✓		✓	✓	26.40	53.02	74.05	16.49	48.00	70.16
SkipPlan [29] ♣	✓				28.85	61.18	**74.98**	15.56	55.64	**70.30**
Ours w/ P^2KG (R=2)	✓				22.60	48.76	53.57	13.90	45.79	55.00
Ours ♣ w/ P^2KG (R=1)	✓				33.34	**61.36**	64.14	20.38	55.54	64.03
Ours ♣ w/ P^2KG (R=2)	✓				**33.38**	60.79	63.89	**21.02**	**56.08**	64.15
PDPP ♣ † [50]	✓			✓	37.20	64.67	66.57	21.48	57.82	65.13
Ours ♣ † w/ P^2KG (R=1)	✓				**38.12**	**64.74**	**67.15**	**24.15**	**59.05**	**66.64**

Models	$T=5$	$T=6$
DDN [9]	3.10	1.20
P^3IV ♣ [55]	7.21	4.40
PDPP ♣ [50]	13.22	7.49
E3P ♣ [49]	8.96	5.76
SkipPlan ♣ [29]	8.55	5.12
Ours (R=2)	8.17	5.32
Ours ♣ (R=1)	**13.25**	8.09
Ours ♣ (R=2)	12.74	**9.23**
PDPP ♣ † [50]	13.45	8.41
Ours ♣ † (R=1)	**14.20**	**9.27**

- R denotes the number of procedure plans used

Results: COIN & NIV

- We compare our model performance with prior works on the datasets CrossTask, COIN, and NIV.

Models	COIN (T=3)			COIN (T=4)			COIN (T=5)		
	$SR\uparrow$	$mAcc\uparrow$	$mIoU\uparrow$	$SR\uparrow$	$mAcc\uparrow$	$mIoU\uparrow$	$SR\uparrow$	$mAcc\uparrow$	$mIoU\uparrow$
Random	< 0.01	< 0.01	2.47	< 0.01	< 0.01	2.32	-	-	-
Retrieval	4.38	17.40	32.06	2.71	14.29	36.97	-	-	-
DDN [9]	13.90	20.19	64.78	11.13	17.71	68.06	-	-	-
P^{3}IV [59]	15.40	21.67	76.31	11.32	18.85	70.53	4.27	10.81	68.81
E3P [53]	19.57	31.42	84.95	13.59	26.72	84.72	-	-	-
PDPP [54]	19.42	43.44	50.03	13.67	42.58	49.84	13.02	**43.36**	50.96
SkipPlan [29]	**23.65**	**47.12**	**78.44**	**16.04**	**43.19**	**77.07**	9.90	38.99	**76.93**
Ours (R=2)	20.25	39.87	51.72	15.63	39.53	53.27	**16.06**	40.72	56.15

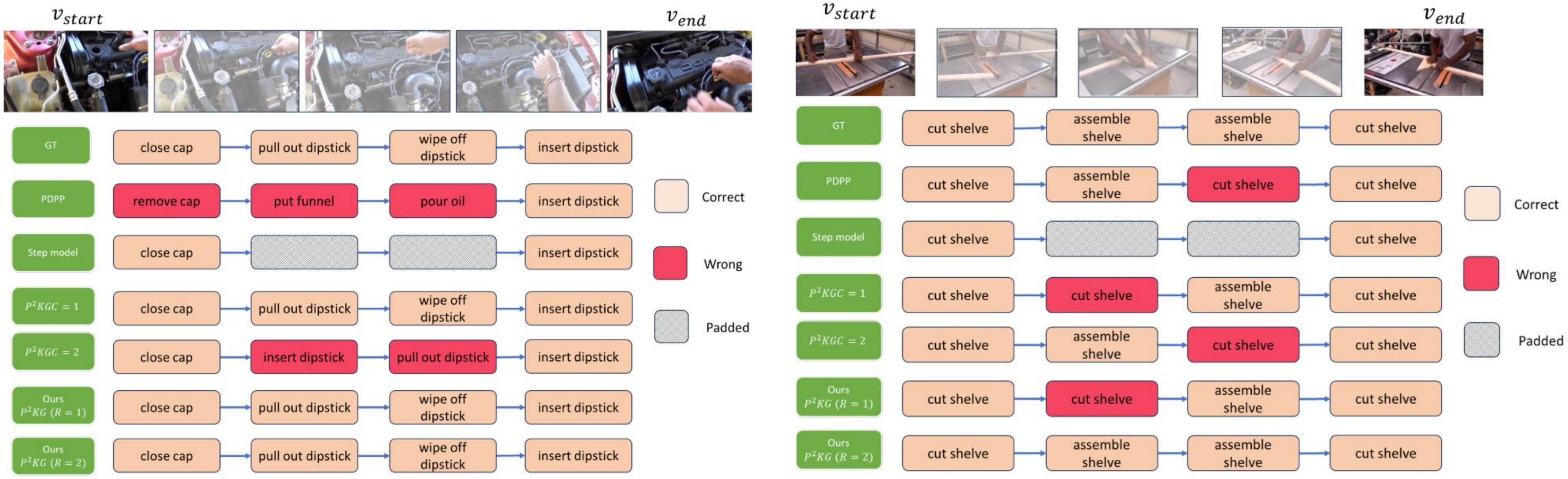
Models	NIV (T=3)			NIV (T=4)		
	$SR\uparrow$	$mAcc\uparrow$	$mIoU\uparrow$	$SR\uparrow$	$mAcc\uparrow$	$mIoU\uparrow$
Random	2.21	4.07	6.09	1.12	2.73	5.84
DDN [9]	18.41	32.54	56.56	15.97	27.09	53.84
Ext-GAIL [7]	22.11	42.20	65.93	19.91	36.31	53.84
P^{3}IV [59]	24.68	49.01	74.29	20.14	38.36	67.29
E3P [53]	**26.05**	**51.24**	75.81	21.37	**41.96**	74.90
PDPP [54]	22.22	39.50	86.66	21.30	39.24	84.96
Ours	24.44	43.46	**86.67**	**22.71**	41.59	**91.49**

Models	NIV (T=5)			NIV (T=6)		
	$SR\uparrow$	$mAcc\uparrow$	$mIoU\uparrow$	$SR\uparrow$	$mAcc\uparrow$	$mIoU\uparrow$
PDPP [54]	18.95	37.26	87.50	14.94	41.02	93.70
Ours (R=2)	**21.58**	**39.79**	**91.66**	**17.53**	**43.62**	**93.75**

- R denotes the number of procedure plans used

Results

- Shown are the qualitative examples of our method.

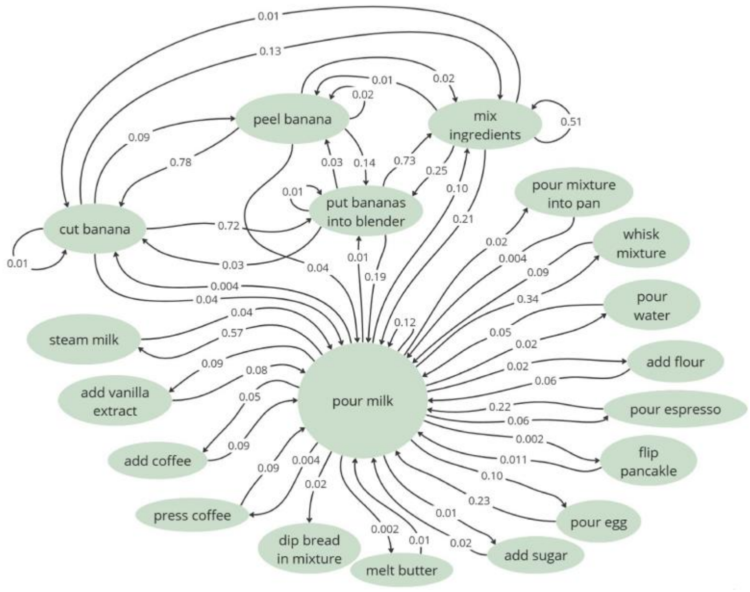


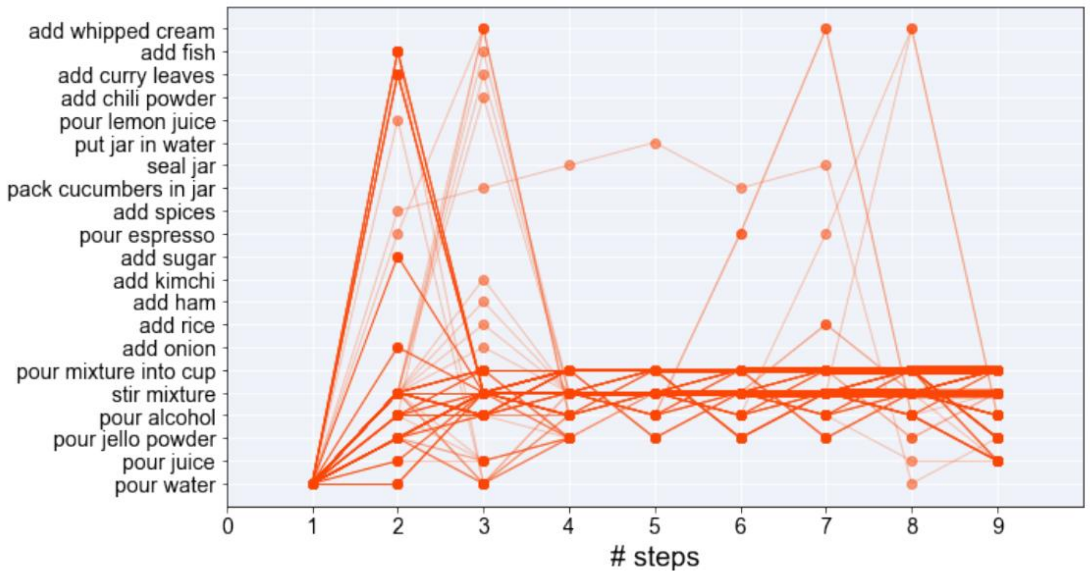
(a) Add Oil to Your Car task (b) Build Simple Floating Shelves task

- Intermediate steps are padded in the step model because it only predicts the start and end actions.

Results: The effect of P²KG

- Procedural knowledge graph effectively encapsulates real-world knowledge of distinct transition probabilities between steps.



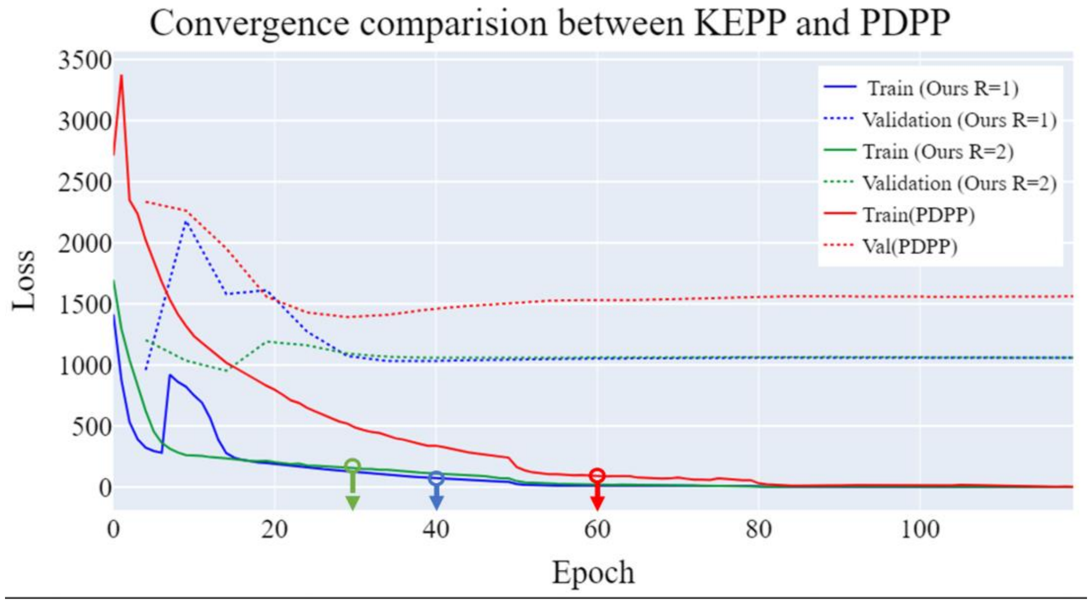


Model	T=3			T=4			T=5			T=6		
	SR	mAcc	mIoU	SR	mAcc	mIoU	SR	mAcc	mIoU	SR	mAcc	mIoU
w.o P^{2}KG conditions †	35.69	63.91	66.04	20.52	57.47	64.39	12.8	53.44	64.01	8.15	50.45	64.13
Ours †	38.12	64.74	67.15	24.15	59.05	66.64	14.20	53.84	65.56	9.27	50.22	65.97
w.o P^{2}KG conditions	31.35	59.51	63.11	18.92	56.20	62.47	12.71	51.29	63.56	8.16	47.63	63.39
Ours	33.38	60.79	63.89	21.02	56.08	64.15	12.74	51.23	63.16	9.23	50.78	65.56

Results: LLMs and Training efficiency

- Utilizing LLMs to enhance action anticipation or planning in other realms

Model (T=6, CrossTask ♣)	SR	mAcc	mIoU
Ours with P^{2}KG (R=1)			
PDPP setting	**9.27**	50.22	**65.97**
Conventional setting	8.09	50.80	**65.39**
One LLM plan recommendation			
PDPP setting (13b)	7.74	50.28	64.05
Conventional setting (13b)	7.21	49.68	63.89
PDPP setting (70b)	8.62	**50.31**	64.34
Conventional setting (70b)	7.81	49.75	64.02
P^{2}KG (R=1) and one LLM plan recommendation			
PDPP setting (13b)	8.81	49.97	65.22
Conventional setting (13b)	8.20	51.46	64.30
PDPP setting (70b)	9.01	50.25	65.57
Conventional setting (70b)	**8.34**	**51.53**	64.96



- Training efficiency comparison between PDPP and our model (KEPP)

Conclusion

- Goal: Generate procedural plans with minimal supervision considering causal constraints in the sequencing of steps and the variability inherent in multiple feasible plans.
- Approach: Infuse procedure planning with comprehensive procedural knowledge, derived from a P^2KG.
- Results:
  - Requires a minimal number of annotations for supervision.
  - Decompose the procedure planning in to two diffusion problems.
  - Experimental evaluations reveal that KEPP attains state-of-the-art results.

SCAN ME