Paper tag: WED-PM-237

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

PDPP: Projected Diffusion for Procedure Planning in Instructional Videos

Hanlin Wang[1] Yilu Wu[1] Sheng Guo[3] Limin Wang[1, 2, ✉]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Shanghai AI Lab, China [3]MYbank, Ant Group, China

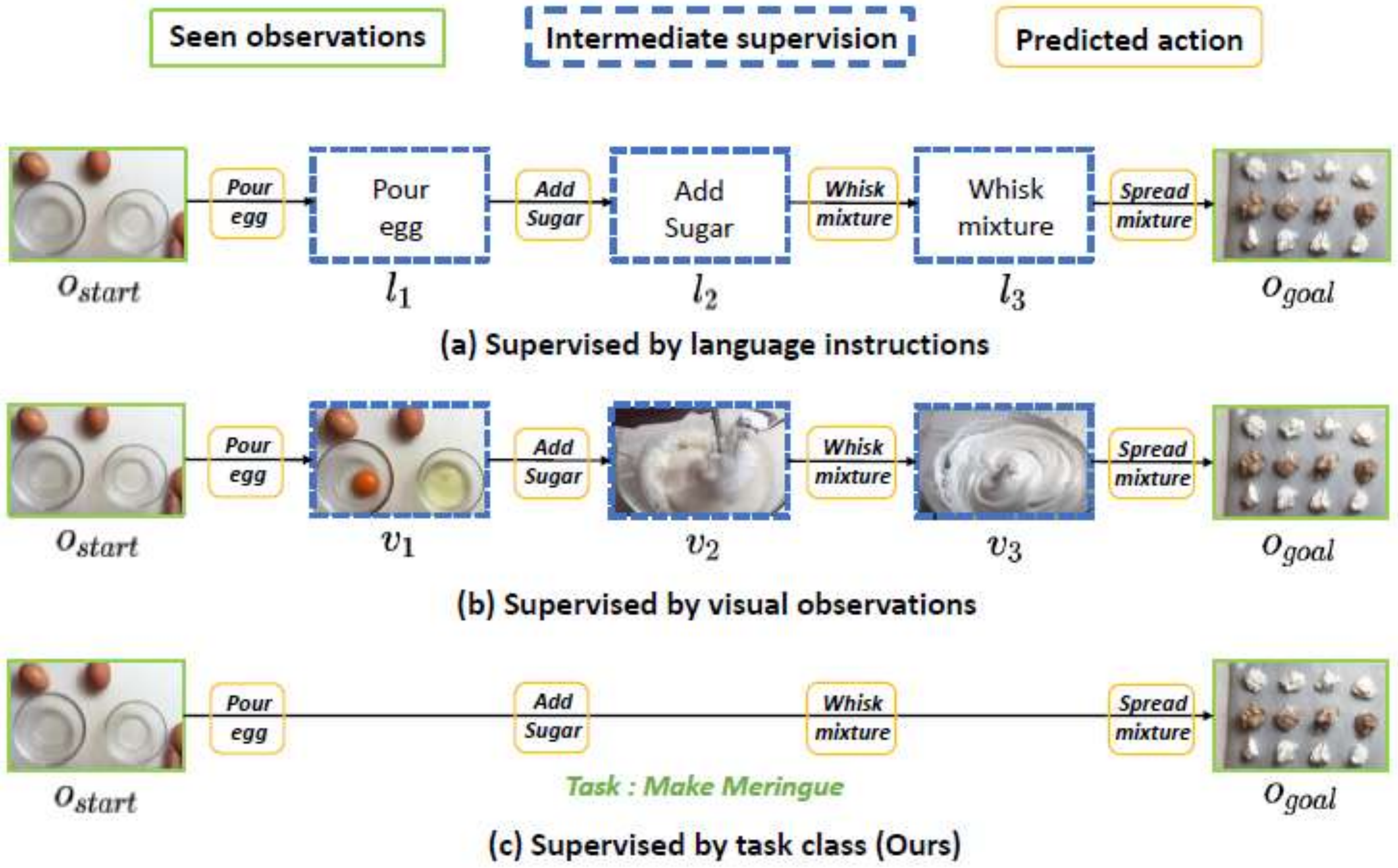
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Quick Preview

Quick Preview

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

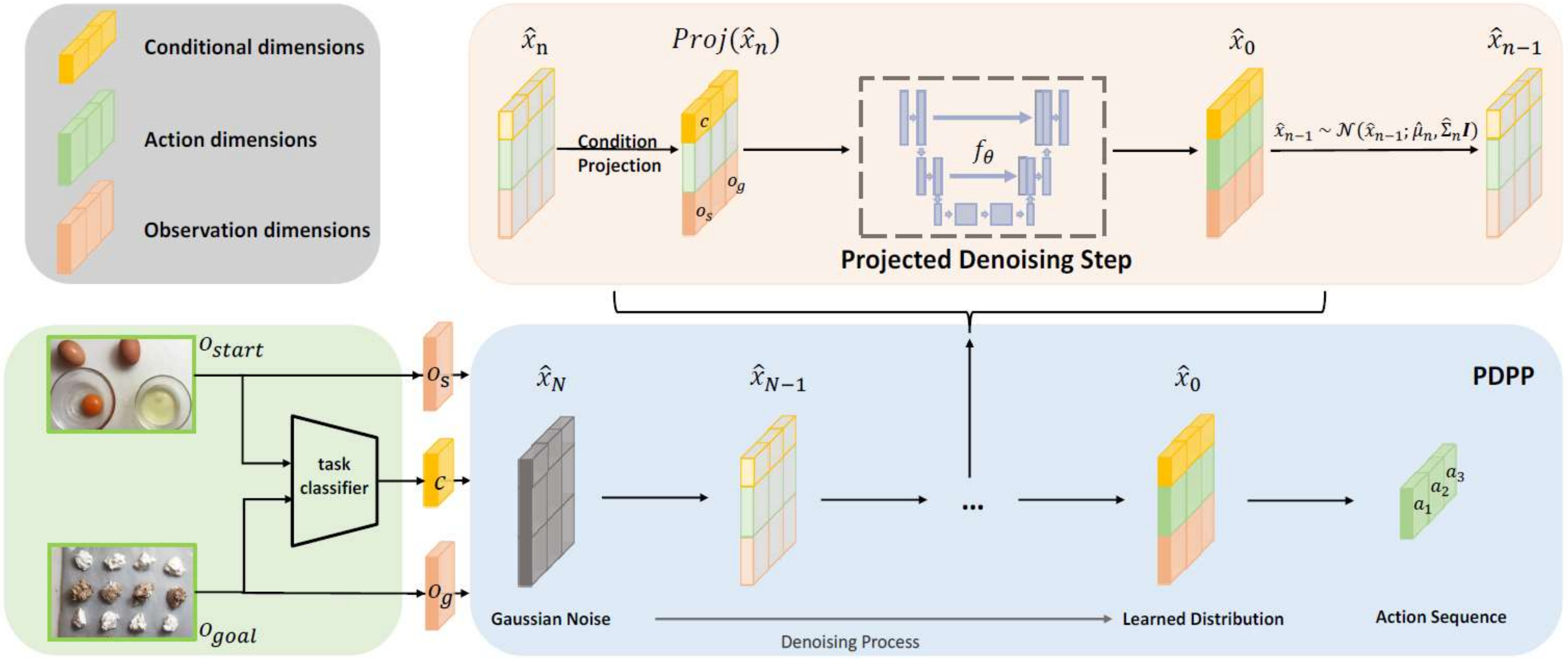
- Procedure planning in instructional videos requires a model to make goal-directed plans, given the current visual observations in unstructured real-life videos.
- We propose a **diffusion-based model, PDPP**, for procedure planning in instructional videos.
- We remove the expensive intermediate supervision, and simply use task labels from instructional videos as supervision instead.



Quick Preview

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- The key insight of PDPP is to **treat this problem as a distribution fitting problem** and model the whole intermediate action sequence distribution.
- To **model the uncertainty** in procedure planning, we propose our PDPP based on **diffusion** models.
- We add the **condition projection operation** into the diffusion process to ensure correct guidance for diffusion.



Quick Preview

- Our PDPP model achieves the state-of-the art performance on three datasets with different scales, even without the task supervision.
- Our PDPP model has an excellent ability to model the uncertainty in procedure planning and can produce both diverse and reasonable plans.

Main Results

Models	$T=3$ SR↑	$T=4$ SR↑	$T=5$ SR↑	$T=6$ SR↑
Retrieval-Based	8.05	3.95	2.40	1.10
DDN [4]	12.18	5.97	3.10	1.20
P³IV [36]	23.34	13.40	7.21	4.40
Ours_{Base}	26.47	15.40	9.37	6.76
Ours_{How}	**37.20**	**21.48**	**13.58**	**8.47**

Models	Supervision	$T=3$ SR↑	$T=3$ mAcc↑	$T=3$ mIoU↑	$T=4$ SR↑	$T=4$ mAcc↑	$T=4$ mIoU↑
Random	-	<0.01	0.94	1.66	<0.01	0.83	1.66
Retrieval-Based	-	8.05	23.30	32.06	3.95	22.22	36.97
WLTDO [10]	-	1.87	21.64	31.70	0.77	17.92	26.43
UAAA [11]	-	2.15	20.21	30.87	0.98	19.86	27.09
UPN [29]	V	2.89	24.39	31.56	1.19	21.59	27.85
DDN [4]	V	12.18	31.29	47.48	5.97	27.10	48.46
Ext-GAILw/o Aug. [2]	V	18.01	43.86	57.16	-	-	-
Ext-GAIL [2]	V	21.27	49.46	61.70	16.41	43.05	60.93
P³IV [36]	L	23.34	49.96	73.89	13.40	44.16	70.01
Ours_{Base}	C	26.47	55.35	58.95	15.40	49.42	56.99
Ours_{How}	C	**37.20**	**64.67**	66.57	**21.48**	**57.82**	65.13

Horizon	Models	Sup.	NIV SR↑	NIV mAcc↑	NIV mIoU↑	COIN SR↑	COIN mAcc↑	COIN mIoU↑
$T=3$	Random	-	2.21	4.07	6.09	<0.01	<0.01	2.47
	Retrieval	-	-	-	-	4.38	17.40	32.06
	DDN [4]	V	18.41	32.54	56.56	13.9	20.19	64.78
	Ext-GAIL [2]	V	22.11	42.20	65.93	-	-	-
	P³IV [36]	L	24.68	49.01	74.29	15.4	21.67	76.31
	Ours	C	**31.25**	**49.26**	57.92	**21.33**	**45.62**	51.82

Horizon	Models	Sup.	NIV SR↑	NIV mAcc↑	NIV mIoU↑	COIN SR↑	COIN mAcc↑	COIN mIoU↑
$T=4$	Random	-	1.12	2.73	5.84	<0.01	<0.01	2.32
	Retrieval	-	-	-	-	2.71	14.29	36.97
	DDN [4]	V	15.97	27.09	53.84	11.13	17.71	68.06
	Ext-GAIL [2]	V	19.91	36.31	53.84	-	-	-
	P³IV [36]	L	20.14	38.36	67.29	11.32	18.85	70.53
	Ours	C	**26.72**	**48.92**	59.04	**14.41**	**44.10**	51.39

Distribution Divergence with GT

Metric↓	Model	T = 3	T=4	T=5	T=6
NLL	Deterministic	**3.57**	4.29	4.70	5.12
	Noise	3.58	4.04	4.45	4.79
	Ours	3.61	**3.85**	**3.77**	**4.06**
KL-Div	Deterministic	**2.99**	3.40	3.54	3.82
	Noise	3.00	3.15	3.30	3.49
	Ours	3.03	**2.96**	**2.62**	**2.76**

Diversity and accuracy of plans

Metric↑	Model	T = 3	T=4	T=5	T=6
SR	Deterministic	**39.03**	21.17	12.59	7.47
	Noise	34.92	18.99	12.04	7.82
	Ours	37.20	**21.48**	**13.58**	**8.47**
ModePrec	Deterministic	**55.60**	**45.65**	35.47	25.24
	Noise	51.04	43.90	34.35	24.51
	Ours	53.14	44.55	**36.30**	**25.61**
ModeRec	Deterministic	34.13	18.35	11.20	6.75
	Noise	**39.42**	25.56	15.67	11.04
	Ours	36.49	**31.10**	**29.45**	**22.68**

Introduction

Introduction

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

- **Previous approaches for procedure planning in instructional videos treat it as a sequence planning problem and focus on predicting each action accurately.**
  - Two-branch autoregressive method to predict the intermediate states and actions step by step: easy to **accumulate errors** during the planning process
  - Transformer-based single branch non-autoregressive model: multiple learning objectives, **complex** training schemes and tedious inference process
  - Require **heavy intermediate** visual or language annotations
- **Modeling the uncertainty in procedure planning is also important.**
  - There might be more than one reasonable plan sequences to transform from the given start state to goal state
  - Eg: change the order of “add sugar” and “add butter” in task “making cake” will not affect the final result

Introduction

Motivation

- **Taking the whole intermediate action sequence distribution as learning objective rather than every discrete action.**
  - Transform the planning problem to a sampling process from the learned distribution
  - Optimize model with a simple MSE loss, which results in less learning objectives and simpler training schemes
  - Use task labels from instructional videos as supervision instead

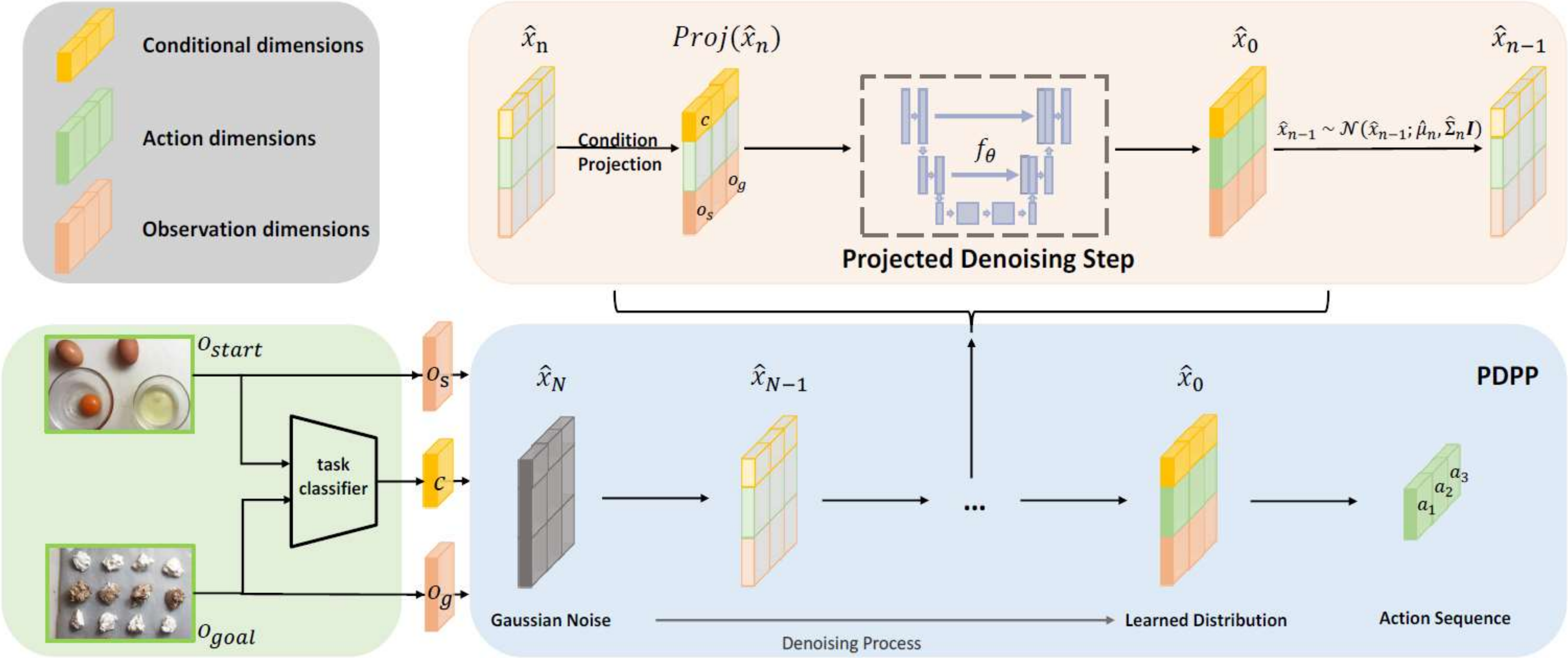
- **Modeling the uncertainty in procedure planning with diffusion model.**
  - Adding randomness to our distribution-fitting process by learning with a diffusion model
  - Convenient to apply conditional diffusion process with the given observations and task class based on diffusion models

Method

Method

Projected Diffusion for Procedure Planning in Instructional Videos（PDPP）

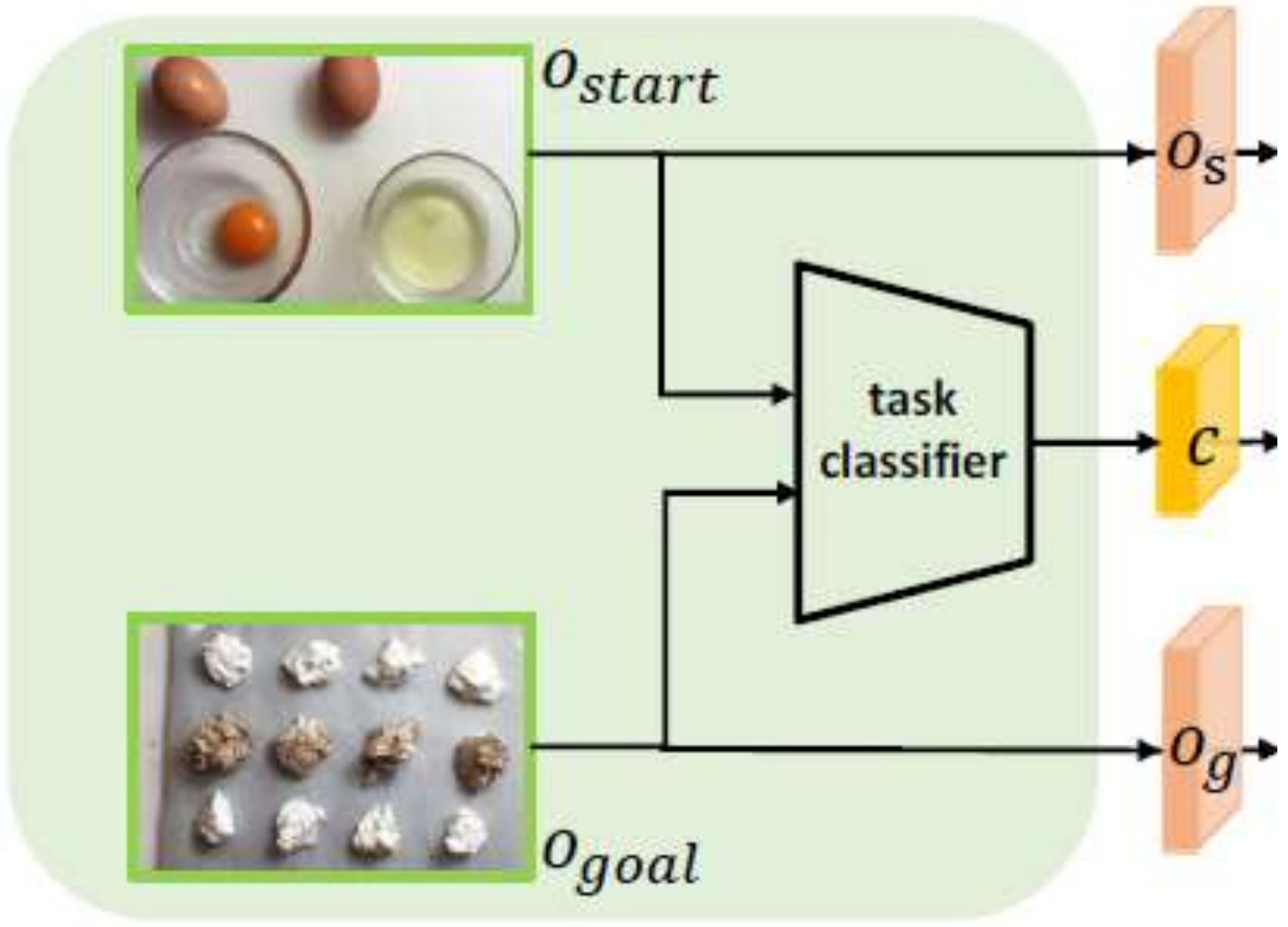
➢ **Conditional action sequence input + Condition projection during learning + Diffusion.**



Method

Step1

➢ **Given the start and goal observations, predict which task category the video is about.**



Make Jello Shots

Make Irish Coffee

...

Build Simple Floating Shelves

- ☐ We implement task classifier with simple MLP models and use the ground truth task labels in instructional videos to supervise the output c.

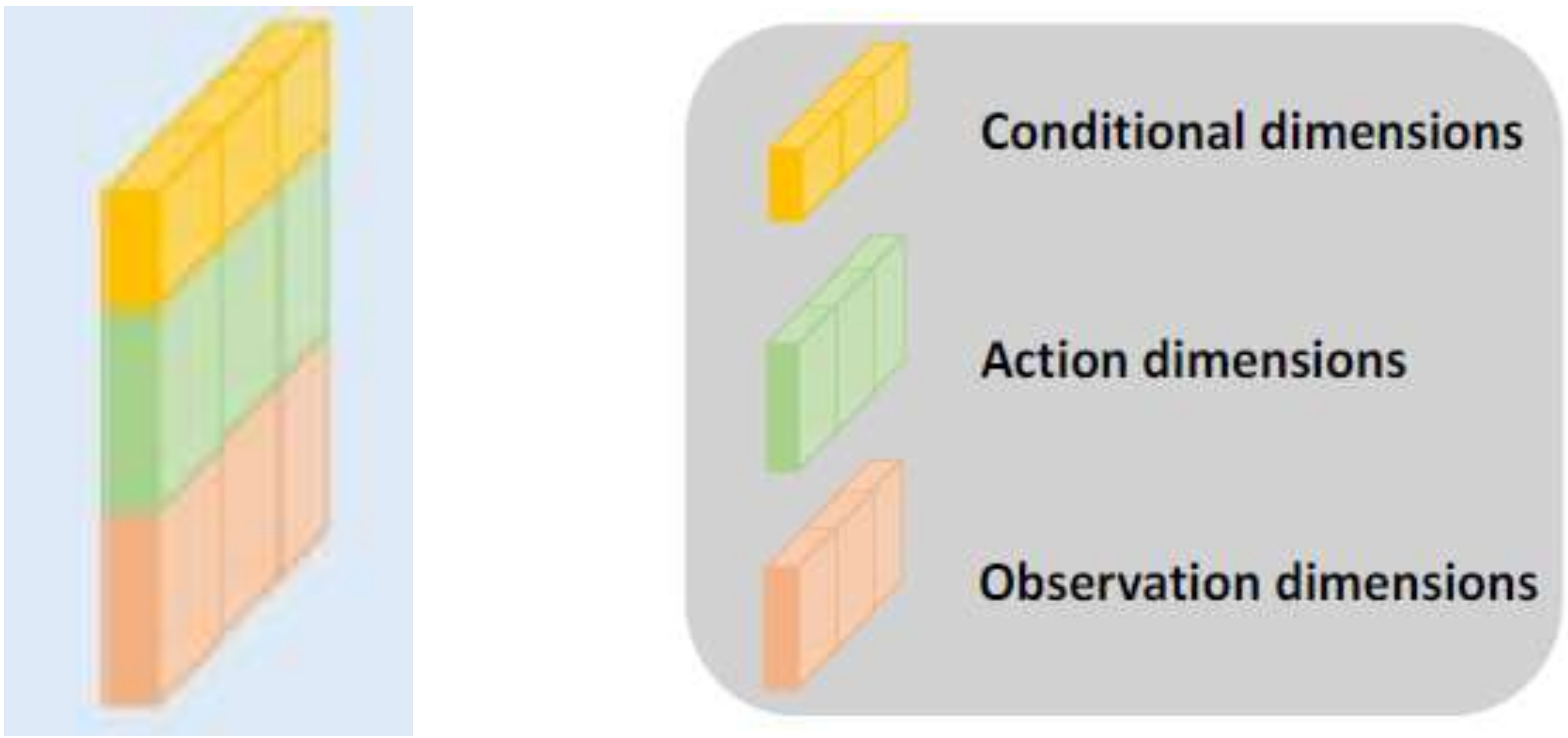
Method

Step2

➢ **Construct conditional action sequence input by concatenate observations, actions(one hot feature) and task class(one hot feature) along the feature dimension. Observations and task class are conditional dimensions.**

$$\begin{bmatrix} c & c & & c & c \\ a_1 & a_2 & \cdots & a_{T-1} & a_T \\ o_s & 0 & & 0 & o_g \end{bmatrix}$$

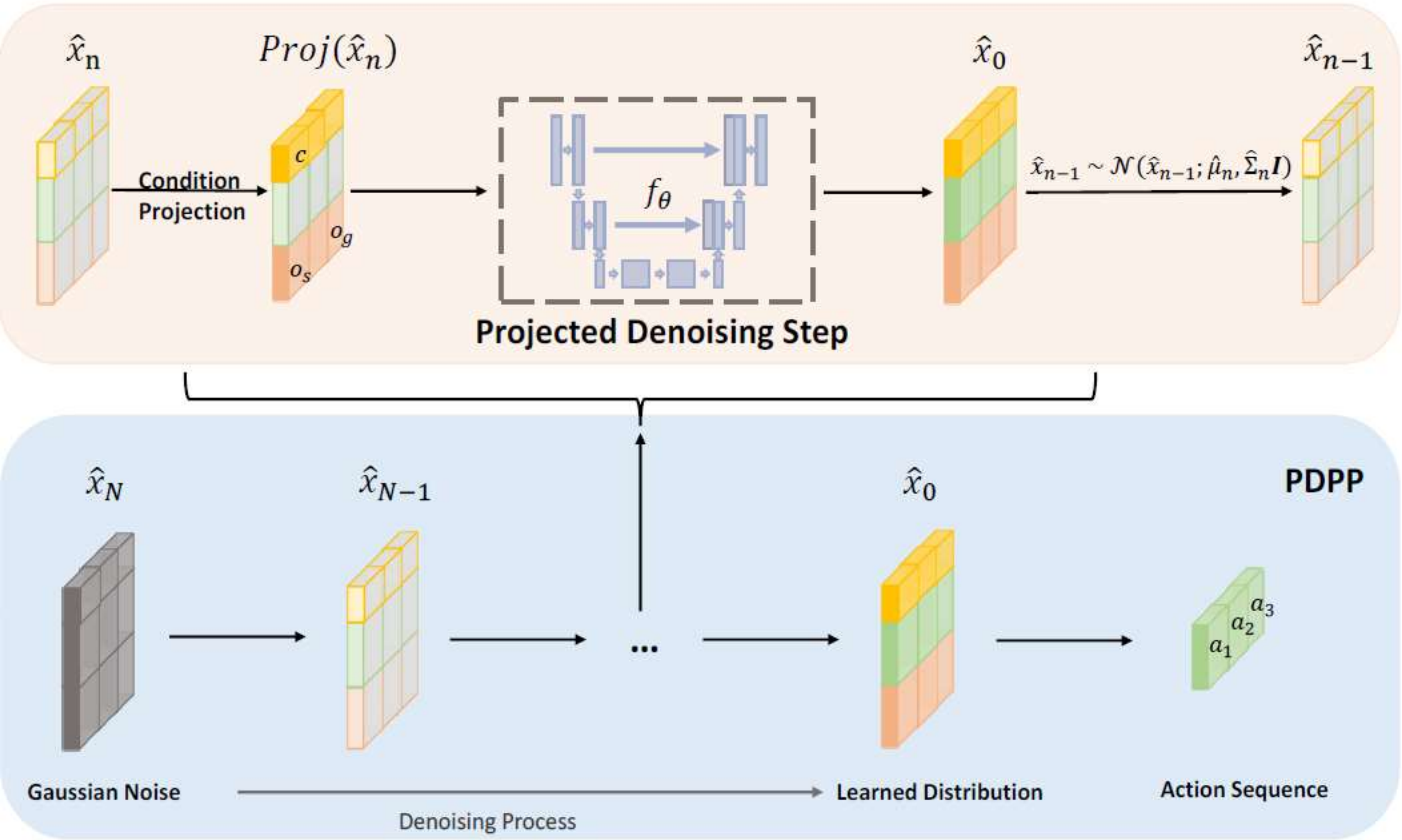
feature dimension



Method

Step3

➢ **Apply diffusion process to the conditional action sequence input. Condition projection is added to both the training and sampling process to ensure correct guidance for diffusion.**



- ☐ We use the basic U-Net as our learnable model for diffusion. Convolution operation along the planning horizon dimension is used for downsample.

- ☐ Learning objective for learnable model f_θ is the initial input x_0 rather than the noise added at each forward diffusion step.

- ☐ Condition projection is implemented by assigning the initial value to observation and task dimensions.

$$\underbrace{\begin{bmatrix} \hat{c}_1 & \hat{c}_2 & & \hat{c}_T \\ \hat{a}_1 & \hat{a}_2 & \cdots & \hat{a}_T \\ \hat{o}_1 & \hat{o}_2 & & \hat{o}_T \end{bmatrix}}_{x} \rightarrow \underbrace{\begin{bmatrix} c & c & & c \\ \hat{a}_1 & \hat{a}_2 & \cdots & \hat{a}_T \\ o_s & 0 & & o_g \end{bmatrix}}_{\text{Proj}(x)}$$

Method

Step3

➢ **Training scheme and sampling process of PDPP.**

Algorithm 1 Training

Input Initial input x_0, total diffusion steps number N, model f_θ, $\{\overline{\alpha}_n\}_{n=1}^N$, weight matrix w

repeat

  $n \sim Uniform(\{1, ..., N\})$

  $\epsilon \sim \mathcal{N}(0, I)$

  $x_n = \sqrt{\overline{\alpha}_n} x_0 + \sqrt{1 - \overline{\alpha}_n}\epsilon$

  $\hat{x}_0 = f_\theta(Proj(x_n), n)$

  Take gradient descent step on

    $\nabla_\theta \|(x_0 - Proj(\hat{x}_0)) * w\|^2$

until converged

- ☐ Weight matrix w is used to assign a bigger weight to a_1 and a_T, since they are actions the most related to the input observations.

Algorithm 2 Inference

Input total diffusion steps number N, model f_θ, $\{\overline{\alpha}_n\}_{n=1}^N$, $\{\beta_n\}_{n=1}^N$

$\hat{x}_N \sim \mathcal{N}(0, I)$

for $n = N, ..., 1$ **do**

  $\hat{x}_0 = f_\theta(Proj(\hat{x}_n), n)$

  if $n > 1$ **then**

    $\hat{\mu}_n = \frac{\sqrt{\overline{\alpha}_{n-1}}\beta_n}{1-\overline{\alpha}_n}\hat{x}_0 + \frac{\sqrt{\alpha_n}(1-\overline{\alpha}_{n-1})}{1-\overline{\alpha}_n}\hat{x}_n$

    $\hat{\Sigma}_n = \frac{1-\overline{\alpha}_{n-1}}{1-\overline{\alpha}_n} \cdot \beta_n$

    $\hat{x}_{n-1} \sim \mathcal{N}(\hat{x}_{n-1}; \hat{\mu}_n, \hat{\Sigma}_n \mathbf{I})$

  end if

end for

return $\hat{x}_0$

Experiments

Experiments

➢ **Evaluation results on CrossTask for procudure planning with prediction horizon T = 3, 4.**

Models	Supervision	$T=3$ SR↑	$T=3$ mAcc↑	$T=3$ mIoU↑	$T=4$ SR↑	$T=4$ mAcc↑	$T=4$ mIoU↑
Random	-	<0.01	0.94	1.66	<0.01	0.83	1.66
Retrieval-Based	-	8.05	23.30	32.06	3.95	22.22	36.97
WLTDO [10]	-	1.87	21.64	31.70	0.77	17.92	26.43
UAAA [11]	-	2.15	20.21	30.87	0.98	19.86	27.09
UPN [29]	V	2.89	24.39	31.56	1.19	21.59	27.85
DDN [4]	V	12.18	31.29	47.48	5.97	27.10	48.46
Ext-GAILw/o Aug. [2]	V	18.01	43.86	57.16	-	-	-
Ext-GAIL [2]	V	21.27	49.46	61.70	16.41	43.05	60.93
P^3IV [36]	L	23.34	49.96	73.89	13.40	44.16	70.01
$Ours_{Base}$	C	26.47	55.35	58.95	15.40	49.42	56.99
$Ours_{How}$	C	**37.20**	**64.67**	66.57	**21.48**	**57.82**	65.13

	Batch size	T = 3	T=4	T=5	T=6
$Ours_{Base}$	1	58.95	56.99	56.32	57.51
	32	68.03	67.14	67.10	**70.48**
	64	**71.46**	**69.64**	**67.39**	69.31
	128	71.01	67.26	64.53	63.19
$Ours_{How}$	1	66.57	65.13	65.32	65.38
	32	75.21	77.07	78.56	78.59
	64	79.74	81.74	**81.73**	**80.88**
	128	**80.50**	**82.32**	81.41	78.64

Evaluation results of mIoU with different batch size on CrossTask.

➢ **Evaluation results on CrossTask for procudure planning with longer planning horizons.**

Models	$T=3$ SR↑	$T=4$ SR↑	$T=5$ SR↑	$T=6$ SR↑
Retrieval-Based	8.05	3.95	2.40	1.10
DDN [4]	12.18	5.97	3.10	1.20
P^3IV [36]	23.34	13.40	7.21	4.40
$Ours_{Base}$	26.47	15.40	9.37	6.76
$Ours_{How}$	**37.20**	**21.48**	**13.58**	**8.47**

Experiments

➢ **Evaluation results on NIV and COIN for procudure planning with prediction horizon T = 3, 4.**

Horizon	Models	Sup.	NIV			COIN		
			SR↑	mAcc↑	mIoU↑	SR↑	mAcc↑	mIoU↑
$T = 3$	Random	-	2.21	4.07	6.09	<0.01	<0.01	2.47
	Retrieval	-	-	-	-	4.38	17.40	32.06
	DDN [4]	V	18.41	32.54	56.56	13.9	20.19	64.78
	Ext-GAIL [2]	V	22.11	42.20	65.93	-	-	-
	P^3IV [36]	L	24.68	49.01	74.29	15.4	21.67	76.31
	Ours	C	**31.25**	**49.26**	57.92	**21.33**	**45.62**	51.82
$T = 4$	Random	-	1.12	2.73	5.84	<0.01	<0.01	2.32
	Retrieval	-	-	-	-	2.71	14.29	36.97
	DDN [4]	V	15.97	27.09	53.84	11.13	17.71	68.06
	Ext-GAIL [2]	V	19.91	36.31	53.84	-	-	-
	P^3IV [36]	L	20.14	38.36	67.29	11.32	18.85	70.53
	Ours	C	**26.72**	**48.92**	59.04	**14.41**	**44.10**	51.39

Experiments

➢ **Ablation study on the role of task supervision.**

	Dataset	w. task sup.			w.o. task sup.		
		SR↑	mAcc↑	mIoU↑	SR↑	mAcc↑	mIoU↑
$T = 3$	CrossTask$_{Base}$	**26.47**	55.35	58.95	22.82	51.56	54.36
	CrossTask$_{How}$	**37.20**	64.67	66.57	35.69	63.91	66.04
	NIV	**31.25**	49.26	57.92	29.41	46.20	56.42
	COIN	**21.33**	45.62	51.82	16.46	36.43	43.50
$T = 4$	CrossTask$_{Base}$	**15.40**	49.42	56.99	14.91	49.55	56.28
	CrossTask$_{How}$	**21.48**	57.82	65.13	20.52	57.47	64.39
	NIV	**26.72**	48.92	59.04	26.72	46.55	59.50
	COIN	**14.41**	44.10	51.39	12.32	35.48	42.75
$T = 5$	CrossTask$_{Base}$	**9.37**	45.93	56.32	8.95	45.77	56.34
	CrossTask$_{How}$	**13.58**	54.05	65.32	12.80	53.44	64.01
$T = 6$	CrossTask$_{Base}$	**6.76**	43.61	57.51	6.06	44.15	57.07
	CrossTask$_{How}$	**8.47**	50.14	65.38	8.15	50.45	64.13

Experiments

➢ Evaluating probabilistic modeling.

Baselines:

- ***Noise:*** remove the diffusion process in PDPP and samples from a random noise with the given observations and task class condition in one shot
- ***Deterministic****:* setting the start distribution as zero thus the model directly predicts a certain result with the given conditions

Metric↓	Model	T = 3	T=4	T=5	T=6
NLL	Deterministic	**3.57**	4.29	4.70	5.12
	Noise	3.58	4.04	4.45	4.79
	Ours	3.61	**3.85**	**3.77**	**4.06**
KL-Div	Deterministic	**2.99**	3.40	3.54	3.82
	Noise	3.00	3.15	3.30	3.49
	Ours	3.03	**2.96**	**2.62**	**2.76**

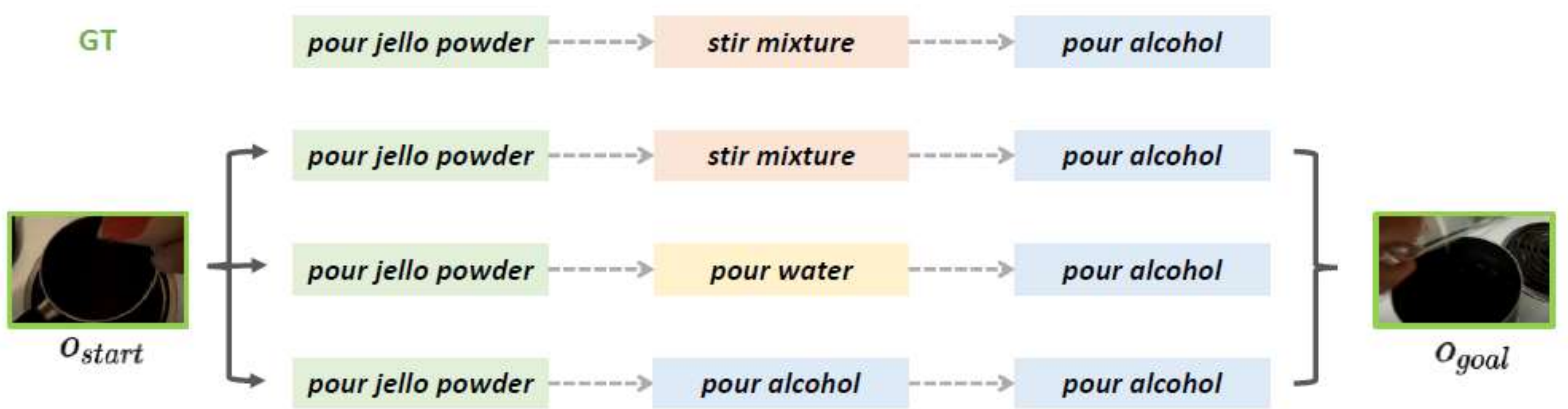
Evaluation results of the plan distributions metrics

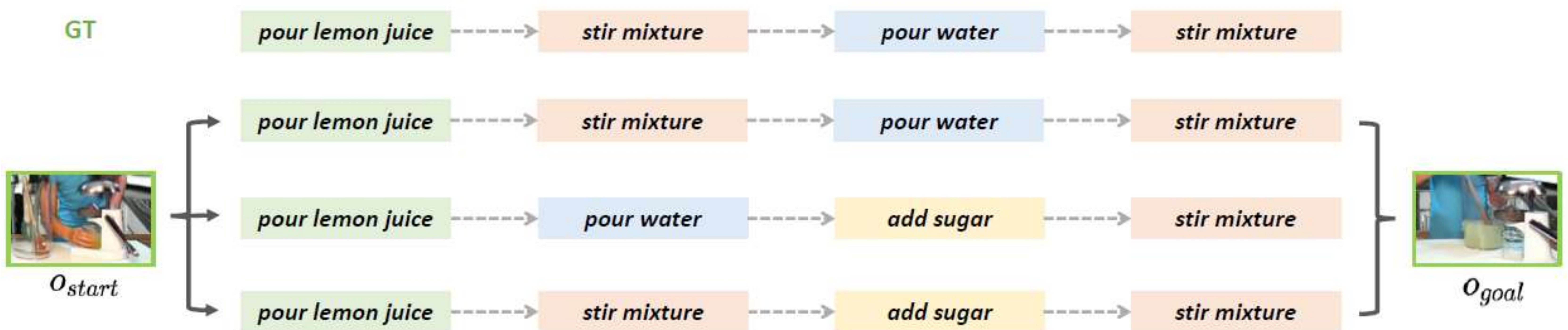
Metric↑	Model	T = 3	T=4	T=5	T=6
SR	Deterministic	**39.03**	21.17	12.59	7.47
	Noise	34.92	18.99	12.04	7.82
	Ours	37.20	**21.48**	**13.58**	**8.47**
ModePrec	Deterministic	**55.60**	**45.65**	35.47	25.24
	Noise	51.04	43.90	34.35	24.51
	Ours	53.14	44.55	**36.30**	**25.61**
ModeRec	Deterministic	34.13	18.35	11.20	6.75
	Noise	**39.42**	25.56	15.67	11.04
	Ours	36.49	**31.10**	**29.45**	**22.68**

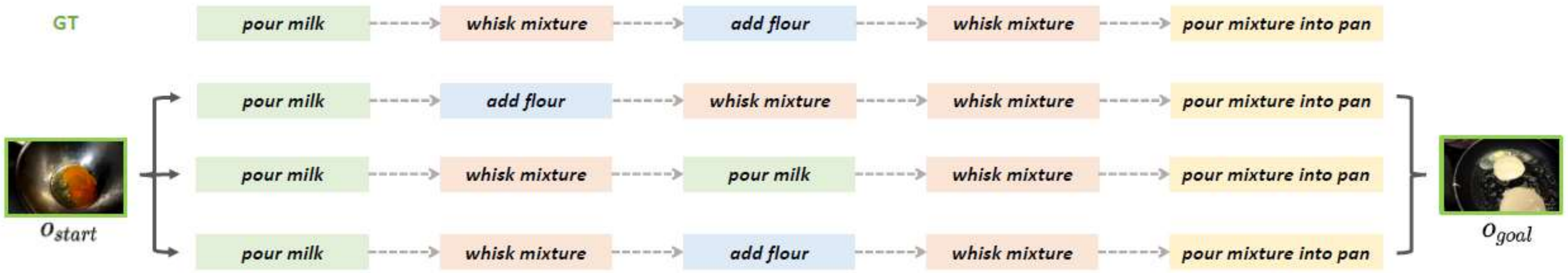
Evaluation results of diversity and accuracy metrics.

Experiments

- **Visualizations for uncertainty modeling.**







Conclusion

Conclusion

- In this work, we cast the procedure planning as a conditional distribution-fitting problem and model the joint distribution of the whole intermediate action sequence as our learning objective, which can be learned with a simple training scheme.

- We introduce an efficient approach for training the procedure planner, which removes the supervision of visual or language features and relies on task supervision instead.

- We propose a novel projected diffusion model (PDPP) to learn the distribution of action sequences and produce all intermediate steps at one shot.

- We evaluate our PDPP on three instructional videos datasets and achieve the state-of-the-art performance across different prediction time horizons.

- Our PDPP model has an excellent ability to model the uncertainty in procedure planning and can produce both diverse and reasonable plans.

Thank you!

Code Link: **https://github.com/MCG-NJU/PDPP**

Contact me: **wwanghanlin06@gmail.com**