Supplemental Material for Skip-Plan: Procedure Planning in Instructional Videos via Condensed Action Space Learning

Zhiheng Li[1], Wenjia Geng[2], Muheng Li[1], Lei Chen[3], Yansong Tang[2]*, Jiwen Lu[1,4], Jie Zhou[1,4]
[1] Department of Automation, Tsinghua University
[2] Shenzhen International Graduate School, Tsinghua University
[3] Beijing University of Science and Technology
[4] Beijing National Research Center for Information Science and Technology
{lizhihan21@,gengwj22@,li-mh20@}mails.tsinghua.edu.cn,chenlei2022@ustb.edu.cn,
{tang.yansong@sz.,lujiwen@,jzhou@}tsinghua.edu.cn

We provide additional information and experimental results in the supplemental material. In Section A, we present the analysis of the error rate distributions on the action chains predicted by different works. It illustrates the behaviour of error accumulation with concrete experimental evidence. Then, we detail the network architecture and loss design at $T = 3$ in Section B. We further demonstrate the robustness and delicacy of our decoupling approach through experiments in Section C.

A. Comparisons of Error Rate Distributions

Error rate analysis is crucial because it reveals how errors are accumulated along action chains. In this section, we compare the error rate distributions along the chains at $T = 4$ predicted by different works, including PlaTe[1], P3IV[2], and our Skip-Plan, as illustrated in Figure S1. Here, the error rate at a timestep t is defined as the number of wrong action predictions at the timestep t divided by the total number of actions at this position, and the relative node position is calculated by $(t-1)/(T-1)$. The error rate distribution is the distribution of the error rates at all timesteps (*e.g.*, the relative node position ranges from 0 to 1). We can summarize two important points from Figure S1. First, we spot the error rate distribution of PlaTe[1] is significantly different from the ones of P3IV[2] and Skip-Plan. The error rate distribution of PlaTe[1] keeps rising until the last action of the chain, but the distributions of P3IV[2] and our Skip-Plan are increasing and then decreasing, reaching a maximum at the middle of the chains. This discrepancy is caused by the network type. PlaTe[1] is an autoregressive network and generates actions one by one. Thus, it accumulates errors continuously from left to right.

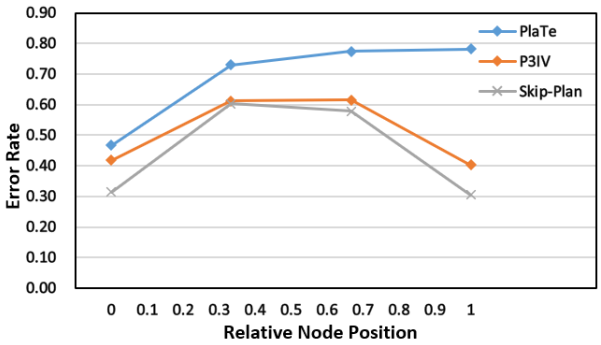

Figure S1. Error rate distributions along the chains predicted by PlaTe[1], P3IV[2], and Skip-Plan at $T = 4$.

In contrast to PlaTe[1], P3IV[2] and our Skip-Plan utilize non-autoregressive transformer decoders to generate whole action sequences in batches. This type of network accumulates the error from two ends of the chain, and the error rate peaks at the middle of the chain. Overall, the average error rate of the non-autoregressive models is lower than the one of the autoregressive models, because the chain length for the error accumulation in the non-autoregressive models is reduced to half compared to the length of the autoregressive models. Second, benefiting from the shared MLP network in the Visual Input Module, the error rates of the first and last actions in our Skip-Plan are the lowest among these works. The decoupling approach without state supervision in Skip-Plan further reduces the compounding error at intermediate actions. Both of them drive our Skip-Plan network to achieve the lowest error rates at all timesteps. Consequently, our Skip-Plan achieves state-of-the-art performance on the CrossTask and COIN benchmarks in procedure planning.

* indicates the corresponding author.

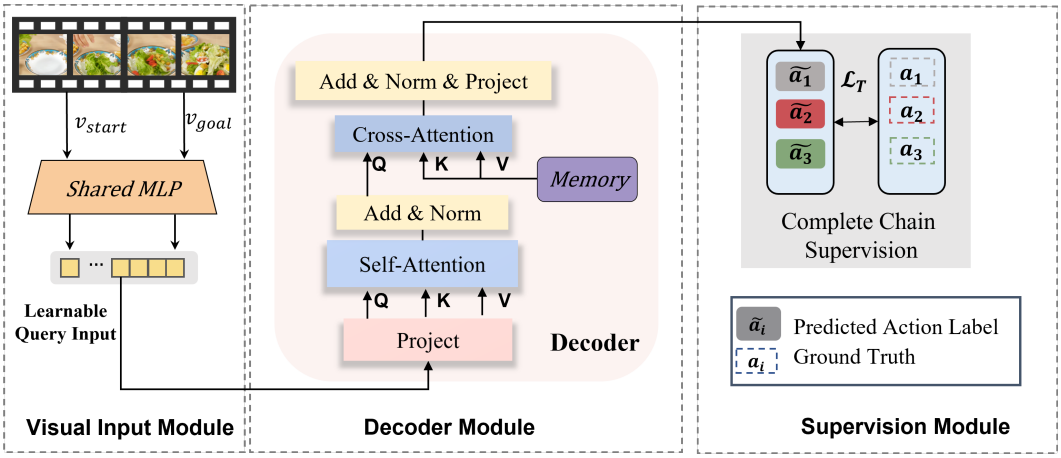

Figure S2. **Our Skip-Plan architecture at T=3.** The network architecture at $T = 3$ is slightly different from Figure 2. It is composed of the Visual Input Module, the Decoder Module, and the Supervision Module, where the Decoder Module and the Supervision Module correspond to the Sub-chain Decoder Module and the Sub-chain Accumulation Module in Figure 2. There is only one decoder in the Decoder Module, which directly outputs the complete chain $\{a_1, a_2, a_3\}$. Without generating sub-chains, there is only one complete chain supervision in the Supervision Module.

B. Network Architecture and Loss for T=3

Our Skip-Plan architecture at $T = 3$ is slightly different from Figure 2. For $T = 3$, the network architecture is composed of the Visual Input Module, the Decoder Module, and the Supervision Module, where the Decoder Module and the Supervision Module just correspond to the Sub-chain Decoder Module and the Sub-chain Accumulation Module in Figure 2 respectively. As illustrated in Figure S2, there is only one decoder in the Decoder Module, which directly outputs the complete chain $\{a_1, a_2, a_3\}$. Without generating sub-chains, no individual sub-chain supervision and sub-chain accumulator are present in the Supervision Module. Therefore, we only have the complete chain supervision, and the loss for $T = 3$ is defined as:

$$\mathcal{L} = FL(a_{1:T}). \tag{S1}$$

C. Reliability of Standalone Sub-chains

To prove our decoupling approach can extract reliable sub-chains at any condition, we further compare the reliability of standalone sub-chains with the reliability of these sub-chains contained within the complete chain at $T = 4/6$. As illustrated in Table S1, the metric results of 'Short' are consistently better than the 'Long' results for all sub-chains at any length. Consequently, our decoupling method can robustly better the prediction results for all long T.

To demonstrate the delicacy of our decoupling method, we try a different decoupling strategy and show how it fails. For example, we choose the sub-chain $\{a_1, a_2, a_3\}$ at $T = 4/5/6$. This sub-chain has the same length as our decoupled sub-chains, but is composed of one reliable initial action and two unreliable intermediate actions, where the actions are all adjacent. In this way, we find the reliability of standalone sub-chains is lower than the one of the sub-chains contained within the original long chains at all long T, illustrated in Table S2. Thus, this type of decoupling approach cannot improve prediction accuracy. The failure of this decoupling method is caused by losing the reliable constraint of the last action. This simple experiment demonstrates our decoupling design is very delicate, robust, and effective.

Table S1. Reliability of standalone sub-chains vs sub-chains contained within original long chains at $T = 4/6$. It validates our decoupling approach can robustly improve the metric results for all long T.

Horizon	Sub-chain	Loss Type	SR	mAcc	mIoU
$T = 4$	$\{a_1, a_2, a_4\}$	Long	15.77	57.58	71.05
		Short	**17.13**	**59.16**	**72.18**
	$\{a_1, a_3, a_4\}$	Long	15.39	57.34	72.29
		Short	**16.88**	**58.93**	**73.06**
$T = 6$	$\{a_1, a_2, a_6\}$	Long	20.87	55.73	69.71
		Short	**23.34**	**58.85**	**71.87**
	$\{a_1, a_3, a_6\}$	Long	15.83	52.76	67.38
		Short	**19.81**	**56.57**	**70.55**
	$\{a_1, a_4, a_6\}$	Long	18.75	54.89	69.10
		Short	**18.95**	**55.21**	**70.06**
	$\{a_1, a_5, a_6\}$	Long	21.07	55.96	70.82
		Short	**22.48**	**57.81**	**72.12**

Table S2. To demonstrate the delicacy of our decoupling method, we try a different decoupling approach and show how it fails. We compare the reliability of the standalone sub-chain $\{a_1, a_2, a_3\}$ vs the sub-chain $\{a_1, a_2, a_3\}$ contained within the original long chain at $T = 4/5/6$. Without the reliable constraint of the last action, the metric results of 'Short' are worse than the 'Long' results at all long T.

Horizon	Loss Type	SR	mAcc	mIoU
$T = 4$	Long	**16.65**	**49.30**	**67.16**
	Short	15.91	48.20	65.93
$T = 5$	Long	**13.10**	**45.67**	**64.46**
	Short	12.82	45.13	62.78
$T = 6$	Long	**11.64**	**44.32**	**62.29**
	Short	11.44	43.30	61.50

References

[1] Jiankai Sun, De-An Huang, Bo Lu, Yun-Hui Liu, Bolei Zhou, and Animesh Garg. Plate: Visually-grounded planning with transformers in procedural tasks. *IEEE Robot. Autom. Lett.*, 7(2):4924–4930, 2022. 1

[2] He Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G Derpanis, Richard P Wildes, and Allan D Jepson. P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In *CVPR*, pages 2938–2948, 2022. 1