Appendix *for* Event-Guided Procedure Planning from Instructional Videos with Text Supervision

An-Lan Wang*, Kun-Yu Lin*, Jia-Run Du, Jingke Meng†, Wei-Shi Zheng†
School of Computer Science and Engineering, Sun Yat-sen University, China
Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{wanganlan, linky5, dujr6}@mail2.sysu.edu.cn, mengjke@gmail.com, wszheng@ieee.org

A. Appendix Overview

Due to the lack of space in the main manuscript, we provide more specific details of our Event-guided Prompting-based Procedure Planning (E3P) in the *Appendix*, organized as follows: Section B provides a more detailed description of our E3P implementation. Section C includes several additional experiments.

B. Implementation Details

Following previous work [A1], we use pre-extracted visual and text features. The dimension of the pre-extracted visual and text features is 512, we use two multi-layer perceptrons (MLP) with shape $[512 \rightarrow 256 \rightarrow 128]$ interspersed with ReLU to embed the original visual and text feature, respectively. For the Event-aware Prompt Generator, the event-information extractor is implemented using a MLP with shape $[256 \rightarrow 64 \rightarrow 128]$, and the event-information aggregator is a Transformer encoder of one self-attention layer with 128-dimensional hidden states. For the Action Relation Mining module, we use two masked self-attention layers followed by a feed-forward network (FFN).

C. Additional Experiments

In addition to the various ablations reported in the main manuscript, we provide some additional experiments to verify the effectiveness of our proposed Event-guided Prompting-based Procedure Planning (E3P).

C.1. Effect of the Event-information Aggregator

For the event-information aggregator, we provide two implementations, ***Concat***, ***Transf*** (*i.e.*, used in the main manuscript):

* indicates equal contribution. † indicates the corresponding author.

Table A1: Effect of Event-information Aggregator for prediction horizon $T \in \{3, 4\}$ on CrossTask dataset. SR and mAcc indicate Success Rate and mean Accuracy, respectively. P3IV is the latest state-of-the-art method.

Model	T = 3		T = 4	
	SR↑	mAcc↑	SR↑	mAcc↑
Concat	25.77	51.32	15.41	47.44
Transf	26.40	53.02	16.49	48.00
P3IV [A1]	23.34	49.96	13.40	44.16

- ***Concat*** first concatenates the prompt representation with the event information and then uses a MLP to project to its original dimension.
- ***Transf*** means using a Transformer encoder of one self-attention layer to process the $T + 1$ Tokens, *i.e.*, T prompt representations and one event information token.

In Table A1, we conduct experiments using different event-information aggregator implementations. "*Transf*" outperforms "*Concat*" in all prediction horizon $T \in \{3, 4\}$ (*i.e.*, 0.63% when $T = 3$ and 1.08% when $T = 4$ in terms of Success Rate). In addition, "*concat*" still achieves state-of-the-art performance, which demonstrates the effectiveness of the proposed event-guided paradigm.

C.2. Analysis of the number of layers used in the Action Relation Mining

In Table A2, we ablate the number of masked self-attention layers used in the Action Relation Mining module (drop rate is 0.2). The results show that using two masked self-attention layers (*i.e.* used in the main manuscript) attains the best performance, *i.e.*, 26.26% when $T = 3$ and 16.49% when $T = 4$ in terms of Success Rate (SR).

Table A2: Quantitative analysis of the number of masked self-attention layers used in the Action Relation Mining module for prediction $T \in \{3, 4\}$ on CrossTask dataset. SR and mAcc indicate Success Rate and mean Accuracy, respectively.

Number of Layers	$T = 3$		$T = 4$	
	SR↑	mAcc↑	SR↑	mAcc↑
1	25.97	52.69	16.10	47.50
2	26.26	52.91	16.49	48.00
3	26.14	52.77	16.15	47.69
4	25.56	52.42	15.69	47.43

Table A3: Comparison to previous state-of-the-art methods using Visual State Supervision for prediction $T = 3$ on CrossTask dataset, in terms of Success Rate (SR), mean Accuracy (mAcc), and mean Intersection over Union (mIoU).

Methods	SR↑	mAcc↑	mIoU↑
baseline	22.86	47.87	70.34
+ event-guided paradigm	25.70	53.19	72.76
P3IV [A1] with visual sup	24.41	45.17	73.83

C.3. Effect of the event-guide paradigm

To verify the effect of the event-guided paradigm in Procedure Planning from instructional videos with Visual Supervision (PPVS), we conduct an experiment that adopts the event-guided paradigm to a variant of P3IV [A1]. In this variant (*i.e.*, baseline), we remove the adversarial strategy and use intermediate visual states as supervision. Then, we insert our proposed Event-guided Prompt Generator (EPG) into this variant (*i.e.*, + event-guided paradigm), but instead of hand-craft prompts, the input to this EPG is learnable queries. The results are shown in A3, by introducing the event-guided paradigm, we attain a significant improvement (*e.g.*, 1.84% in terms of Success Rate), outperforming P3IV [A1] with visual state supervision (*i.e.*, P3IV with visual sup). These consistent results demonstrate the effectiveness of our proposed event-guided paradigm for PPVS.

References

[A1] 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*, 2022. 1, 2