ViterbiPlanNet: Injecting Procedural Knowledge via Differentiable Viterbi for Planning in Instructional Videos

Highlight

Luigi Seminara

Department of Mathematics and Computer Science

University of Catania

Last Year PhD Student

Davide Moltisanti
University of Bath

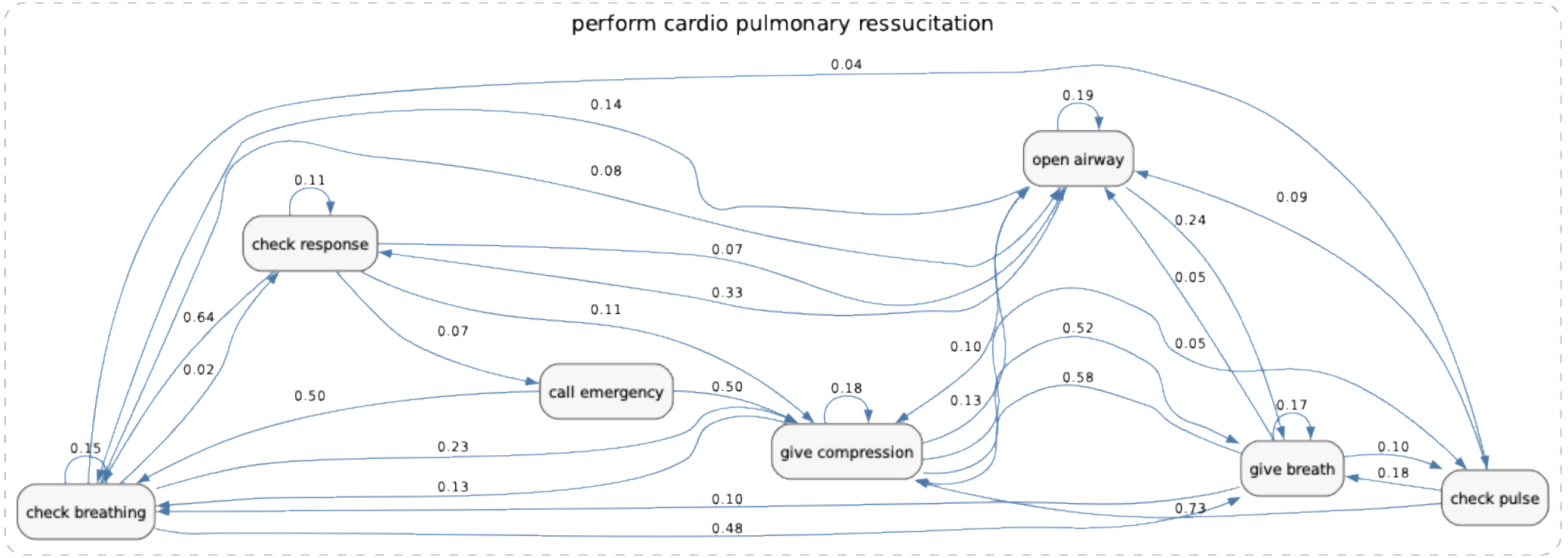
Antonino Furnari
University of Catania

@luseminara.bsky.social

@Gigii_Gii

luigi-seminara

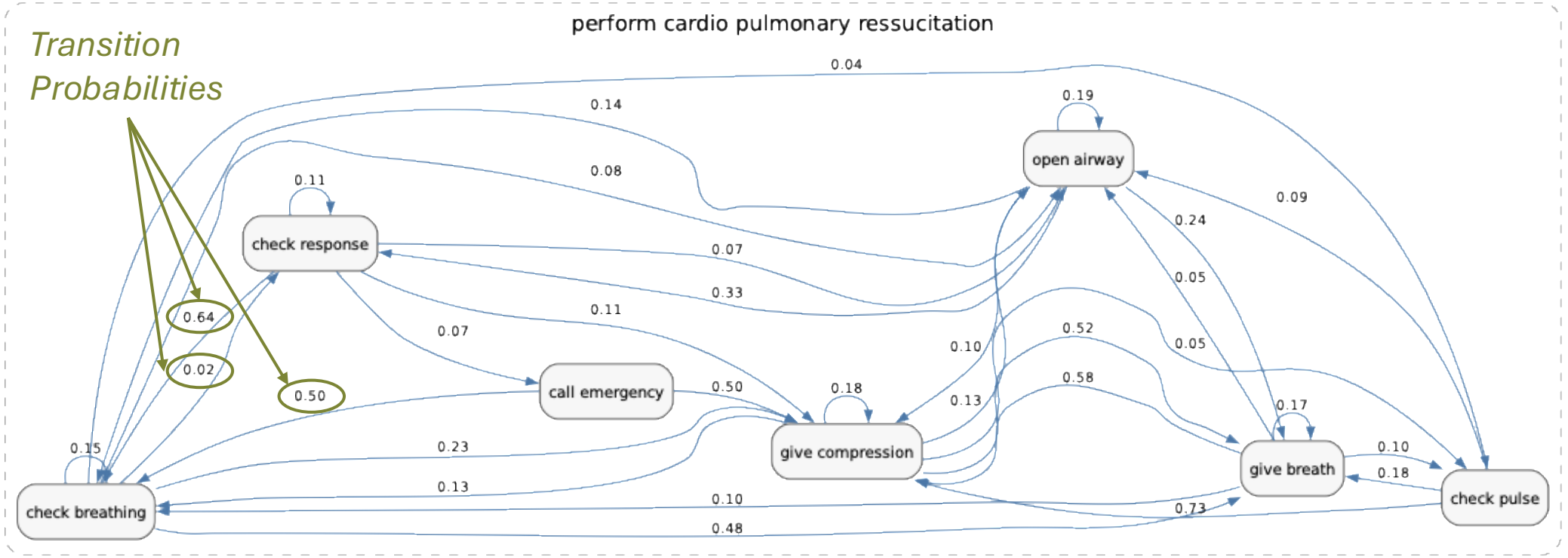
What is a Procedural Knowledge Graph?



Example from NIV dataset

- *Alayrac, Jean-Baptiste, et al. "Unsupervised learning from narrated instruction videos." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.*

What is a Procedural Knowledge Graph?



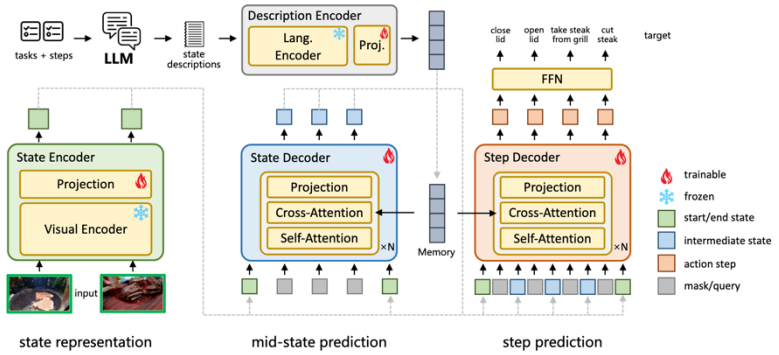
Example from NIV dataset

- *Alayrac, Jean-Baptiste, et al. "Unsupervised learning from narrated instruction videos." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.*

Prior Works

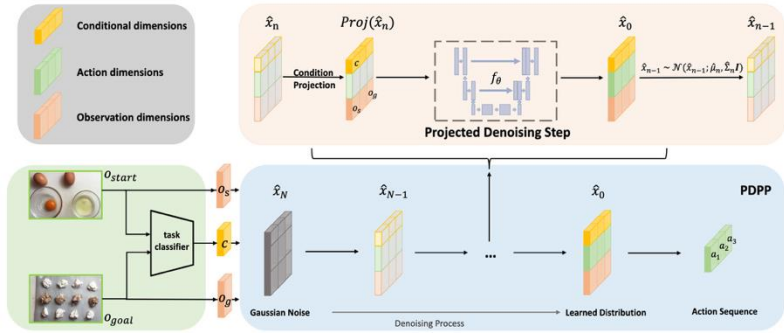
SCHEMA: State CHangEs MAtter for Procedure Planning in Instructional Videos

Yulei Niu[1] **Wenliang Guo**[1] **Long Chen**[2] **Xudong Lin**[1] **Shih-Fu Chang**[1]
[1]Columbia University [2]The Hong Kong University of Science and Technology
yn.yuleiniu@gmail.com



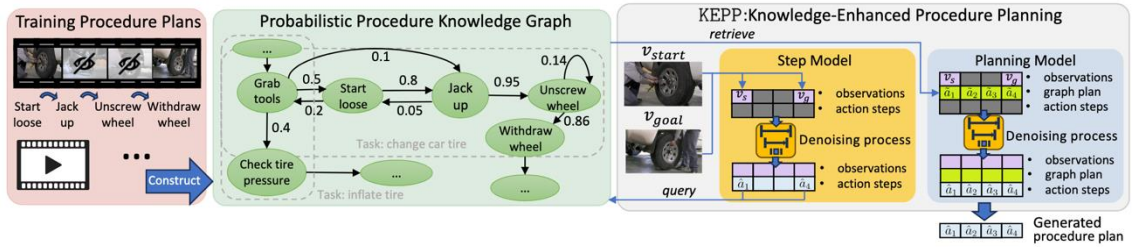
PDPP: Projected Diffusion for Procedure Planning in Instructional Videos

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



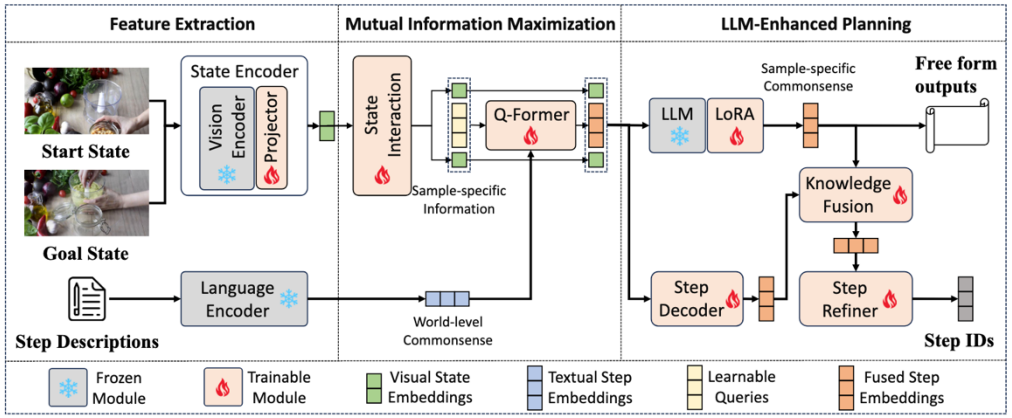
Why Not Use Your Textbook? Knowledge-Enhanced Procedure Planning of Instructional Videos

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



PlanLLM: Video Procedure Planning with Refinable Large Language Models

Dejie Yang[1], **Zijing Zhao**[1], **YangLiu**[1,2*]
[1] Wangxuan Institute of Computer Technology, Peking University
[2] State Key Laboratory of General Artificial Intelligence, Peking University
{ydj,zijingzhao}@stu.pku.edu.cn, yangliu@pku.edu.cn

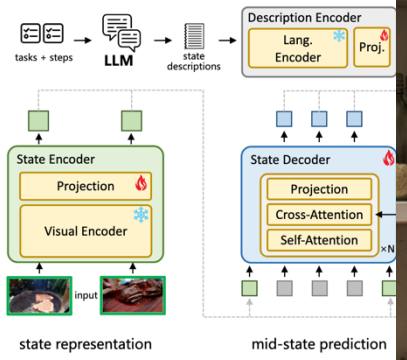


Prior Works

CVPR JUNE 3-7, 2026 DENVER COLORADO

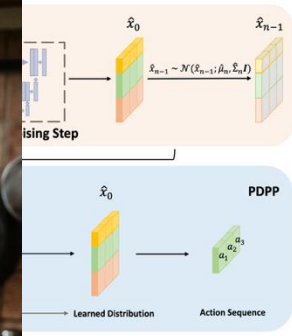
SCHEMA: State CHangEs MAtter for Procedure Planning in Instructional Videos

Yulei Niu[1] **Wenliang Guo**[1] **Long Chen**[2] **Xudong Lin**[1] **Shih-Fu Chang**[1]
[1]Columbia University [2]The Hong Kong Uni
yn.yuleiniu@gmail.com



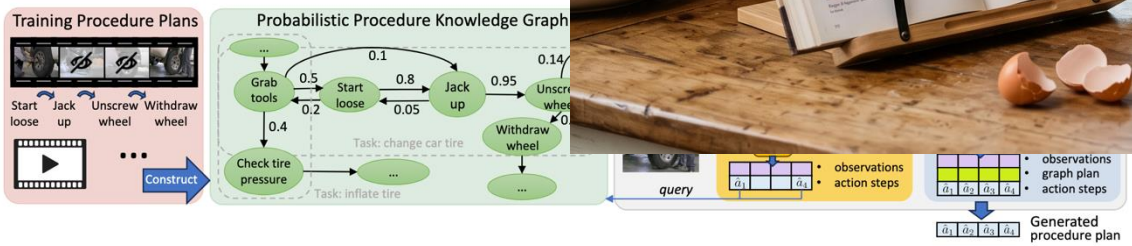
PDPP: Projected Diffusion for Procedure Planning in Instructional Videos

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



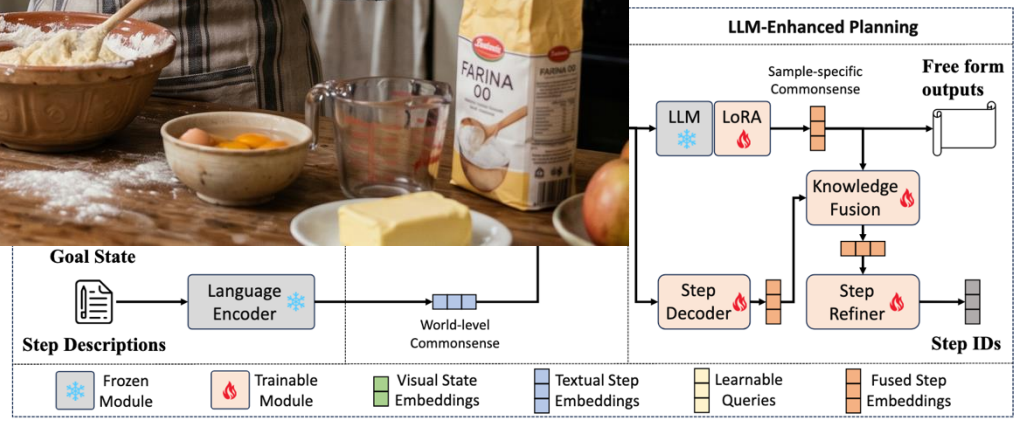
Why Not Use Your Textbook? Knowle… Instructio…

Kumaranage Ravindu Yasas Nagasinghe[1]
Martin Renqiang Min[2] Daniel H…
[1]Mohamed bin Zayed University of Artifici…
[3] University of Auckland, [4]W…

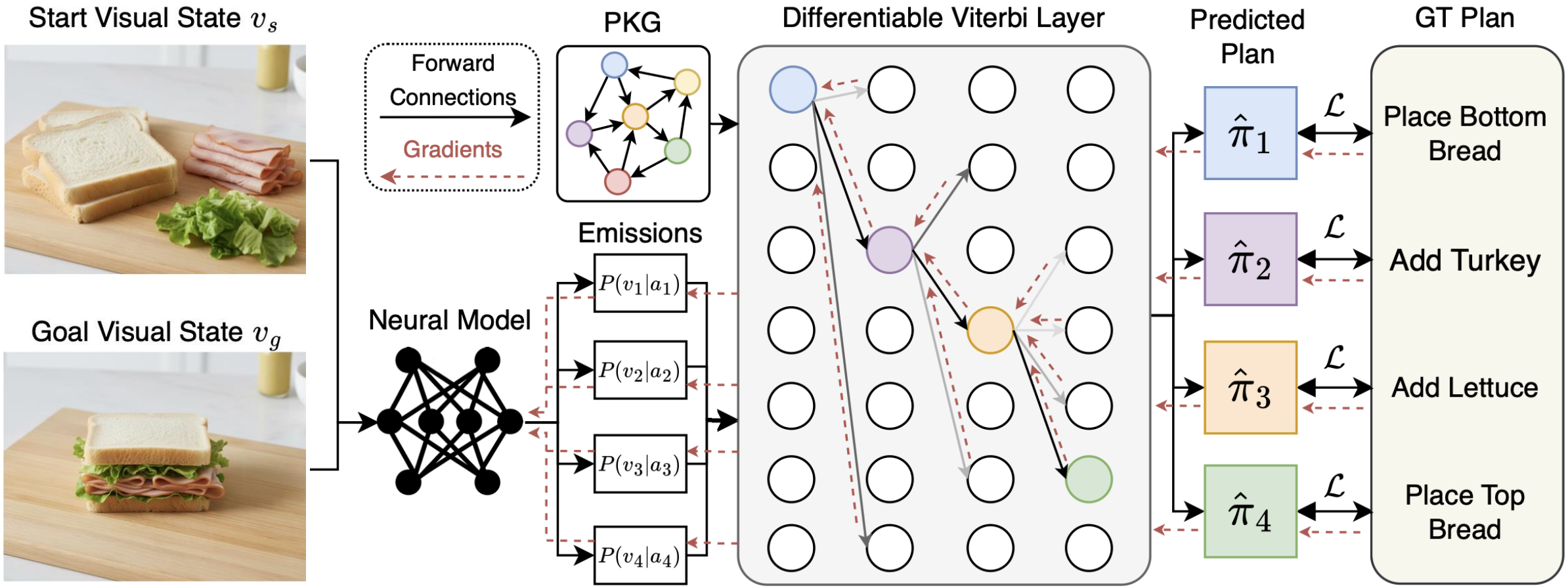


…ble Large Language Models

…gLiu[1,2*]
…eking University
…ce, Peking University
…@pku.edu.cn

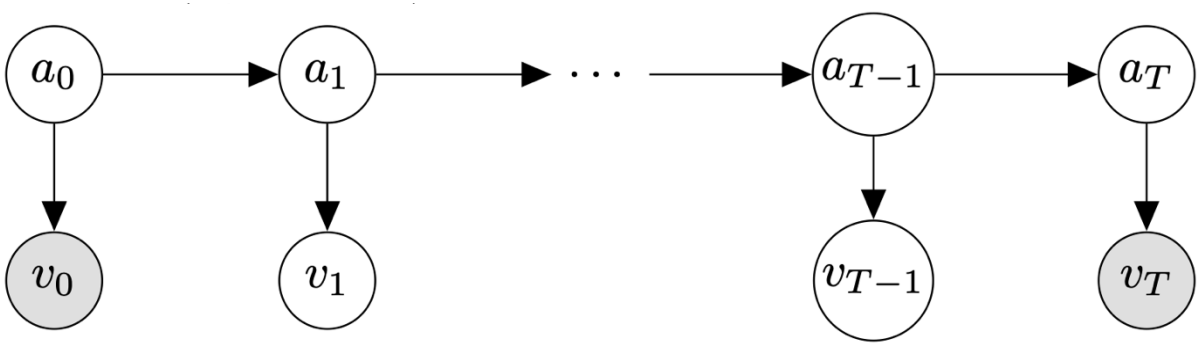


Using Procedural Graphs



Cognitive Offloading to the Graph Available both at Training and Inference

Probabilistic Framework



$$P(a_{0:T}, v_{0:T}) = P(a_0)P(v_0|a_0)\prod_{t=1}^{T} P(a_t|a_{t-1})P(v_t|a_t)$$

$$\pi^* = \underset{\pi=a_{1:T}\in\mathcal{K}^T}{\arg\max} \prod_{t=1}^{T} \underbrace{P(a_t|a_{t-1})}_{\text{Transition}} \underbrace{P(v_t|a_t)}_{\text{Emission}}$$

Viterbi Algorithm

$$\text{S-max}(\mathbf{x}) = \log\left(\sum_{k=1}^{N}\exp(x_k - m)\right) + m$$

$$\text{S-argmax}(\mathbf{x})_k = \frac{\exp(x_k - m)}{\sum_{j=1}^{N}\exp(x_j - m)}$$

Differentiable Relaxations

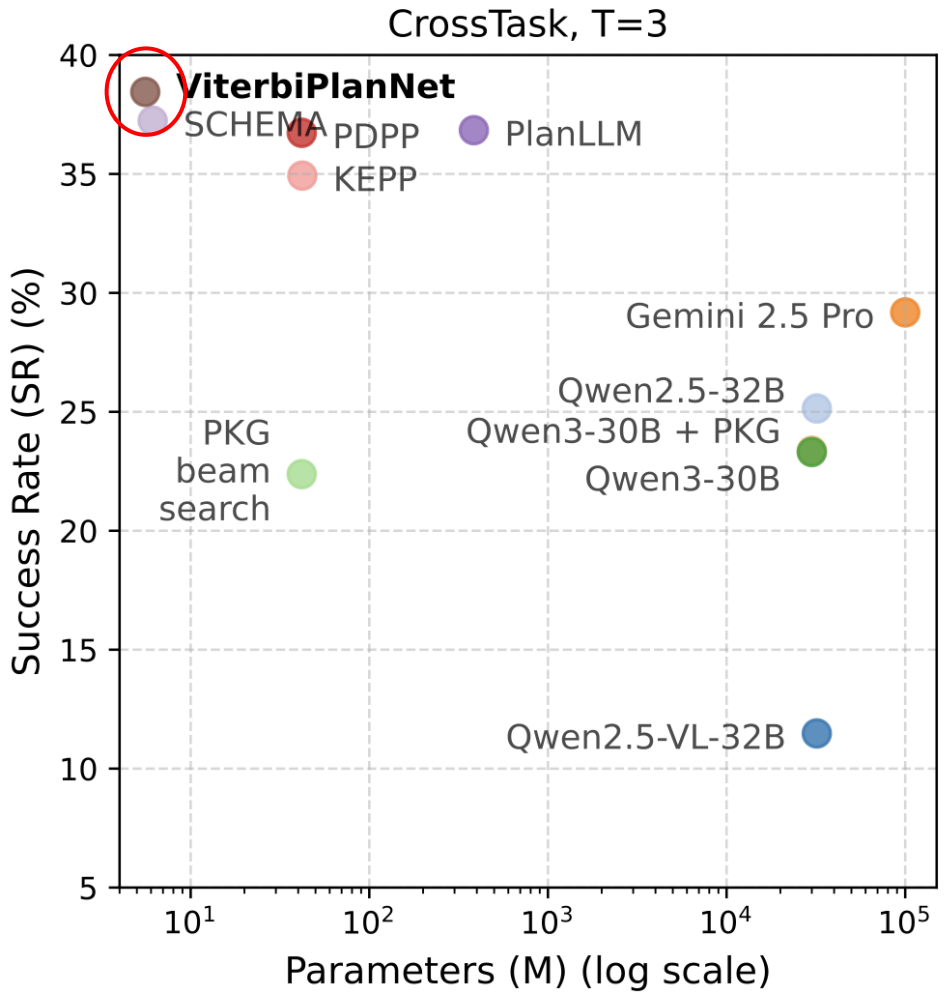
Algorithm 1 Differentiable Viterbi Layer (DVL)

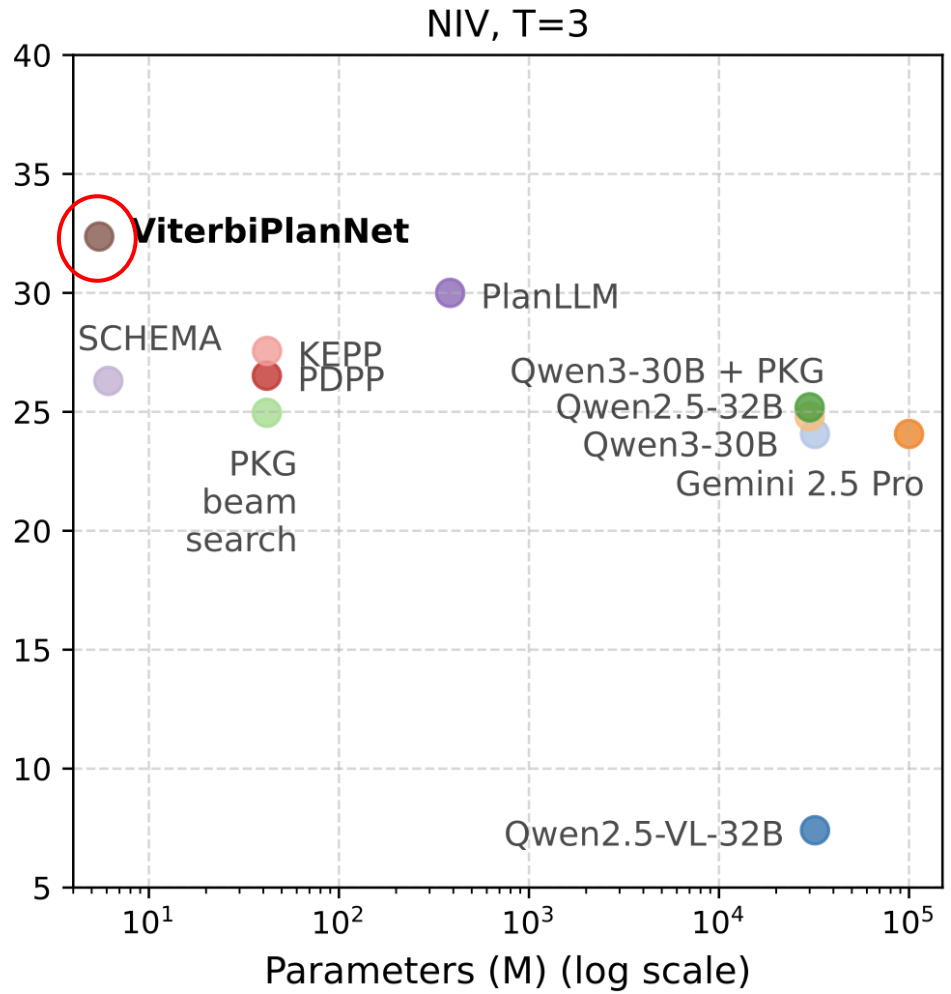
Input: Emission probabilities $b \in [0,1]^{T\times N}$, transition probabilities $\omega(i,j)$ from PKG
Output: Soft plan distribution $\tilde{\pi} \in [0,1]^{T\times N}$

// 1. Initialization (t=1)
for $j = 1$ **to** N **do**
  $\delta_1(j) \leftarrow b[1,j]$ ▷ Initial state scores
end for
// 2. Forward Pass: Differentiable Recursion ($t > 1$)
for $t = 2$ **to** T **do**
  for $j = 1$ **to** N **do**
    // Compute predecessor scores
    for $i = 1$ **to** N **do**
      $s_{i\to j}^{(t)} \leftarrow \delta_{t-1}(i)\cdot\omega(i,j)$
    end for
    // Update state scores using smooth max
    $\delta_t(j) \leftarrow b[t,j]\cdot\text{S-max}(\{s_{i\to j}^{(t)}\}_{i=1}^{N})$
    // Compute soft backpointer distribution
    $\psi_t(j,\cdot) \leftarrow \text{S-argmax}(\{s_{i\to j}^{(t)}\}_{i=1}^{N})$
  end for
end for
// 3. Backward Pass: Recursive Composition
$\tilde{\pi}_T \leftarrow \text{S-argmax}(\delta_T)$ ▷ Distribution at the final horizon
for $t = T-1$ **down to** 1 **do**
  $\tilde{\pi}_t \leftarrow \sum_{j=1}^{N}\tilde{\pi}_{t+1}(j)\cdot\psi_{t+1}(j,\cdot)$ ▷ Recursive backtrace
end for
return $\tilde{\pi}$

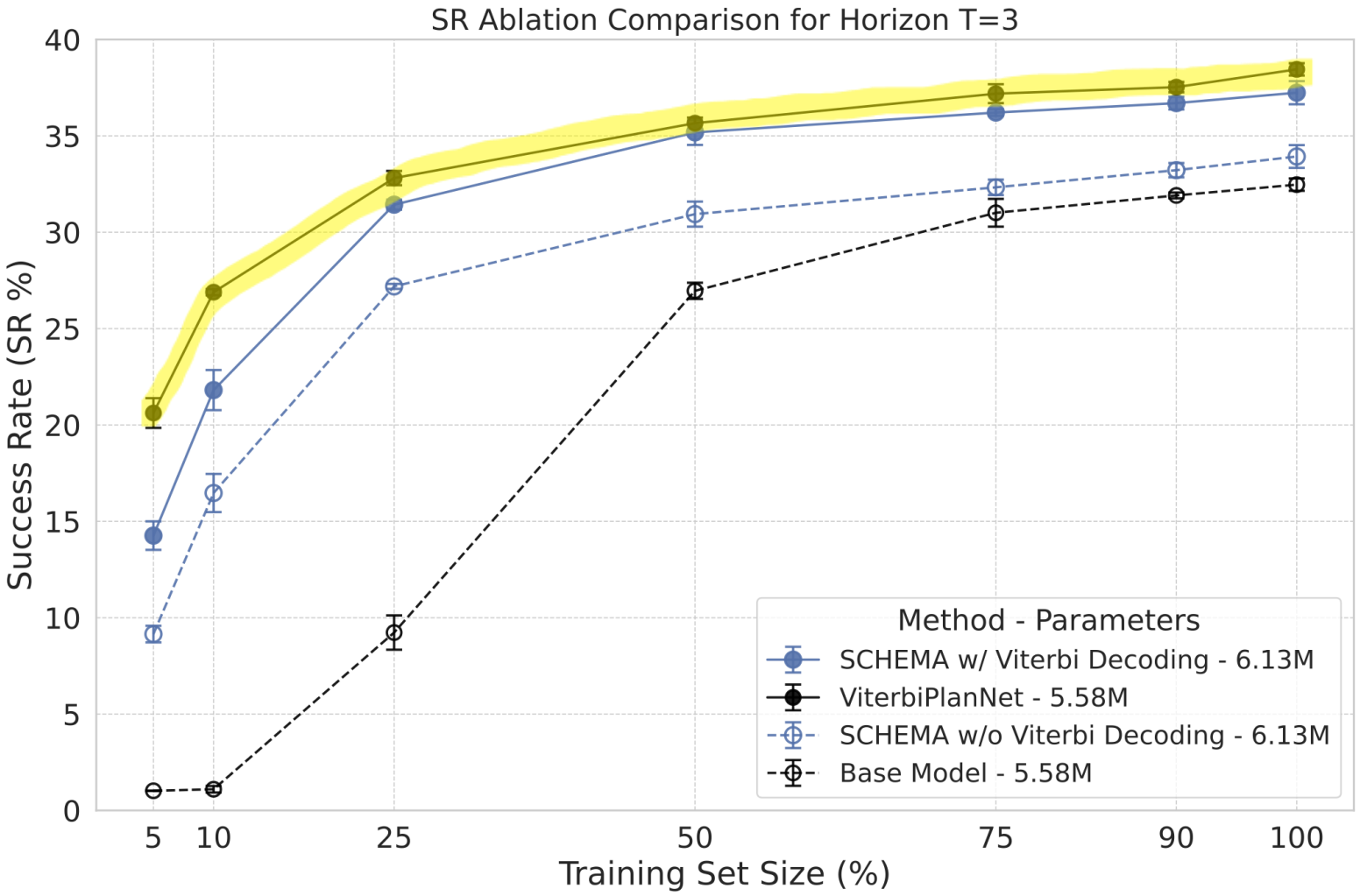
Allows Gradients to Pass

Better Performance with Fewer Parameters



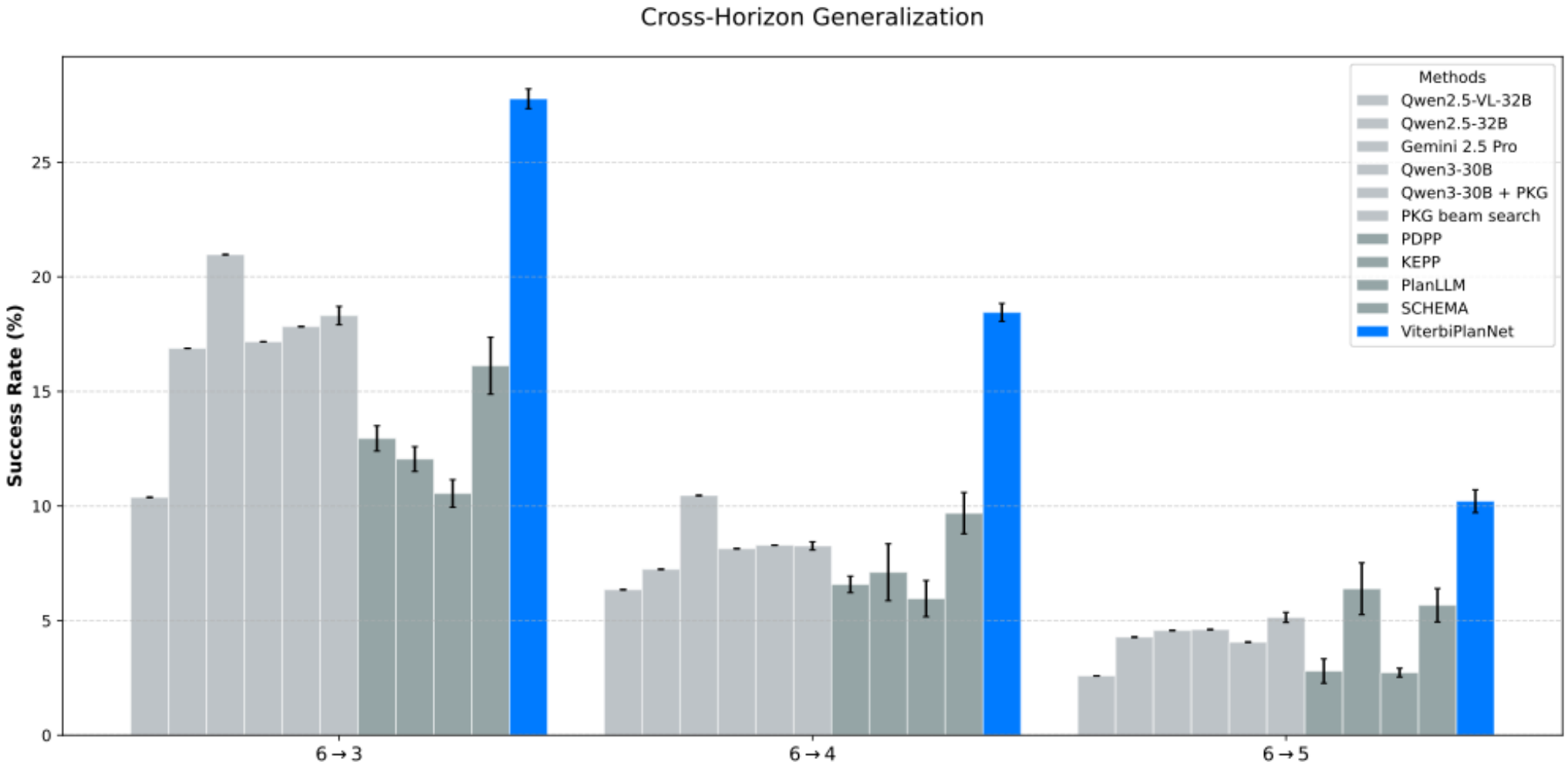


Sample Efficiency



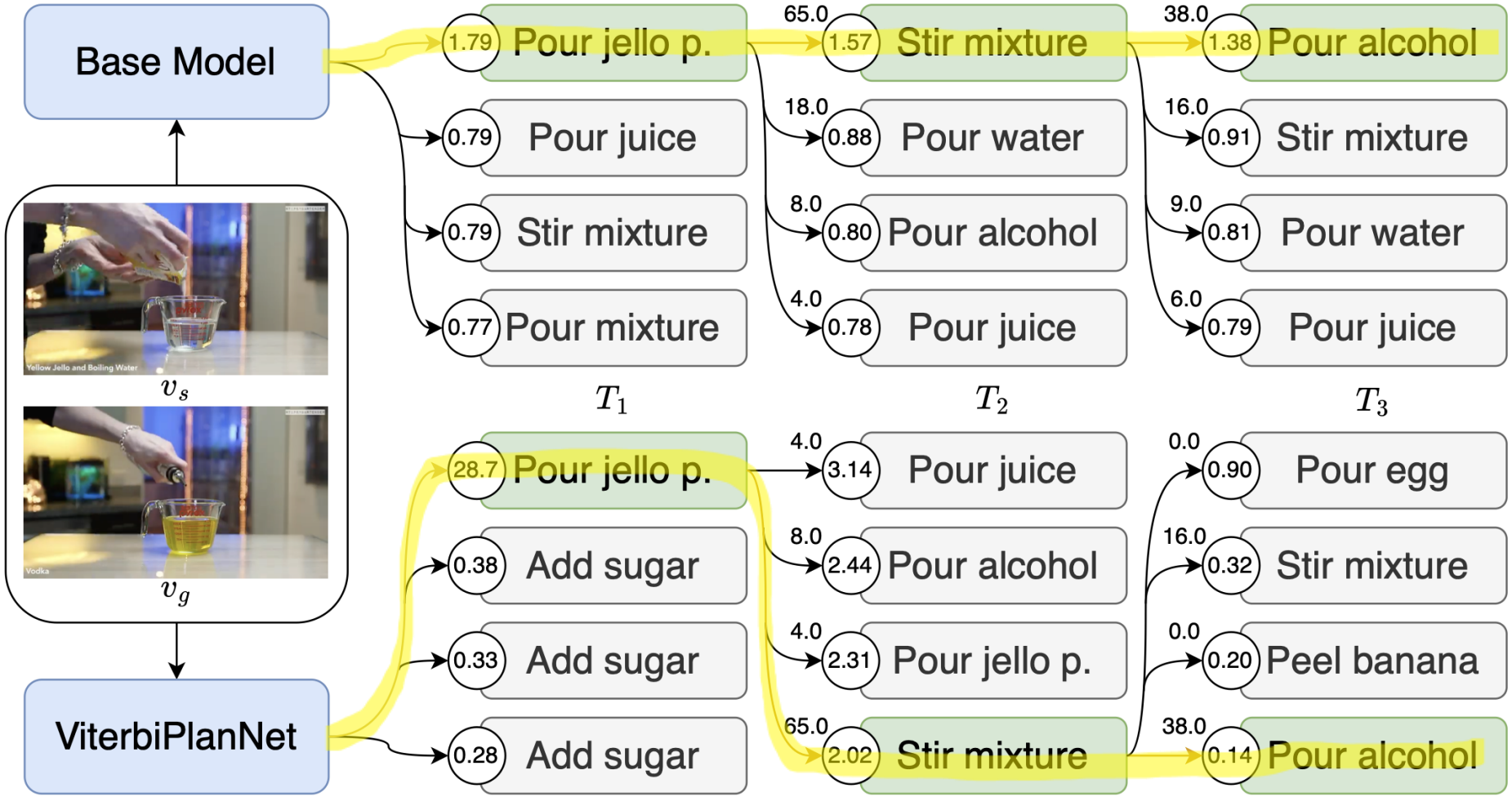
Good performance with fewer training examples.

Cross-Horizon Generalization



Generalizes when trained on a longer horizon and tested on a shorter one.

Qualitative Example

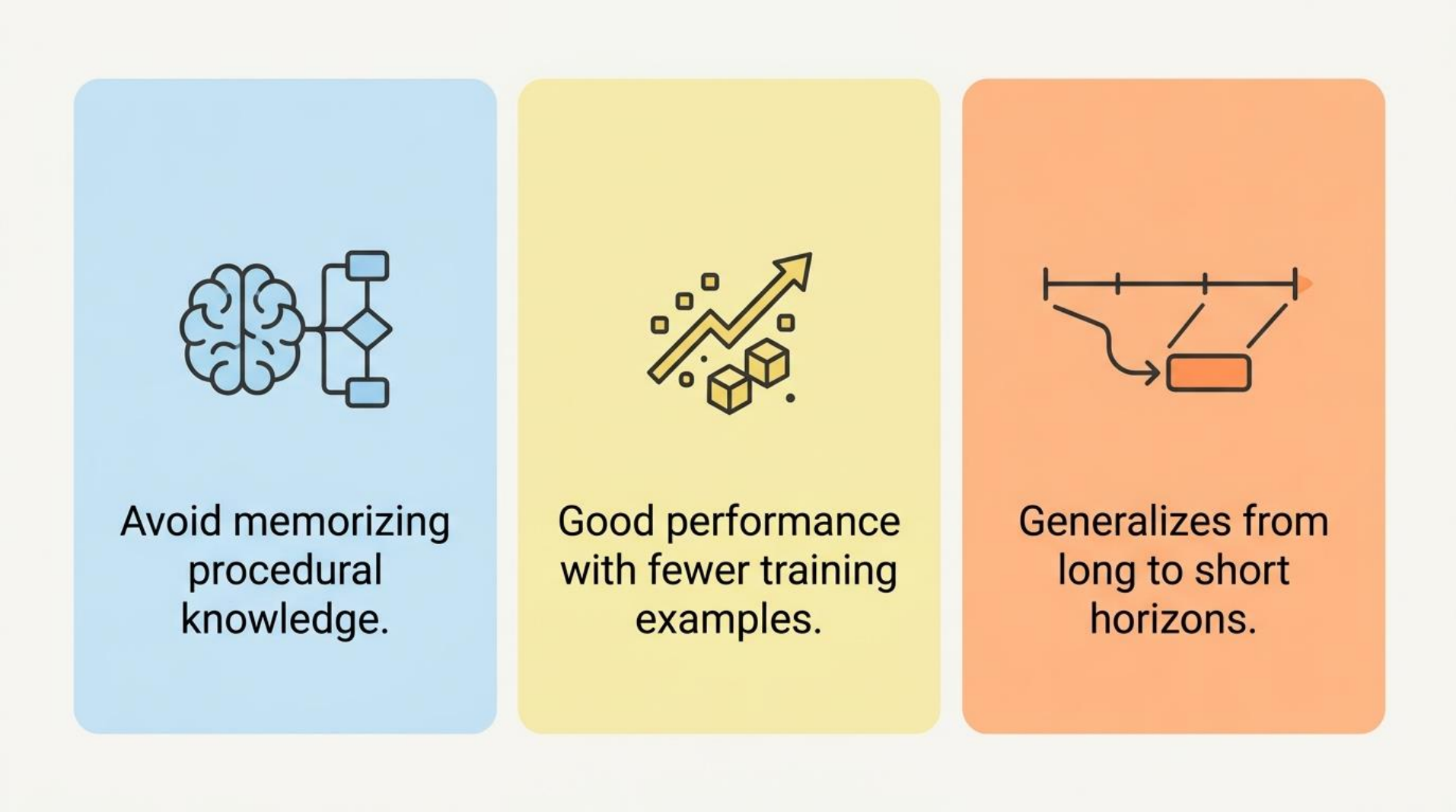


The base model learns to memorize the PKG.

ViterbiPlanNet just has to get a reasonable sorting, relying on the graph for correction.

Easier Cognitive Task

Summary


Avoid memorizing procedural knowledge.
Good performance with fewer training examples.
Generalizes from long to short horizons.

ViterbiPlanNet: Injecting Procedural Knowledge via Differentiable Viterbi for Planning in Instructional Videos

Highlight

Luigi Seminara

Department of Mathematics and Computer Science

University of Catania

Last Year PhD Student

Davide Moltisanti
University of Bath

Antonino Furnari
University of Catania

@luseminara.bsky.social

@Gigii_Gii

luigi-seminara