Contrastive Energy Fields for Inference-Time Procedure Planning in Instructional Videos

Mohamed Afham[1,3] Christoph Reich[1,2,3,4] Oliver Hahn[1] Daniel Cremers[2,3,4] Stefan Roth[1,3,5]

1 TECHNISCHE UNIVERSITÄT DARMSTADT 2 3 ZUSE SCHOOL ELIZA 4 Munich Center for Machine Learning 5 hessian.AI

Project Page

TL;DR

CEFITO (**C**ontrastive **E**nergy **F**ields for **I**nference-**T**ime **O**ptimization) casts procedure planning as **energy minimization**: a predictor maps the initial state and a candidate plan to a predicted goal state; the distance to the observed goal is the plan's energy. Contrastive learning trains this predictor: the ground-truth plan is pulled towards the observed goal, wrong plans are pushed away. At inference time, a predicted task label restricts the search to task-relevant actions, and the lowest-energy plan is returned.



Energy of a candidate plan



Inference-time search on the predicted task's action space

Introduction

Procedure planning [2] predicts the action sequence $a_{1:\mathrm{T}}$ that turns a visual *initial state* into a visual *goal state*, both given as video frames.

Motivation

- Existing planners predict the action sequence with feed-forward, diffusion, or knowledge-graph models [3, 5, 7], treating **every action as plausible**, **without any task constraints**
- Humans instead infer the set of possible actions and *optimize* over it to reach the goal [1]; the same inference-time optimization has shown effective in robotics and navigation

Goal

- Learn an action-conditioned energy field with a *contrastive* objective and no intermediate-state supervision
- Plan by *optimizing* over that field at inference time, restricted to the actions of the predicted task

Method: Learning the Energy Field

A predictor P_θ maps the initial-state latent x_s (frozen S3D) and a CLIP-encoded candidate plan $a_{1:\mathrm{T}}$ to a predicted goal latent. Its distance to the observed goal x_g is the **energy**:

$$E(a_{1:\mathrm{T}}) \;=\; \big\| P_\theta(x_s, a_{1:\mathrm{T}}) - x_g \big\|_2$$

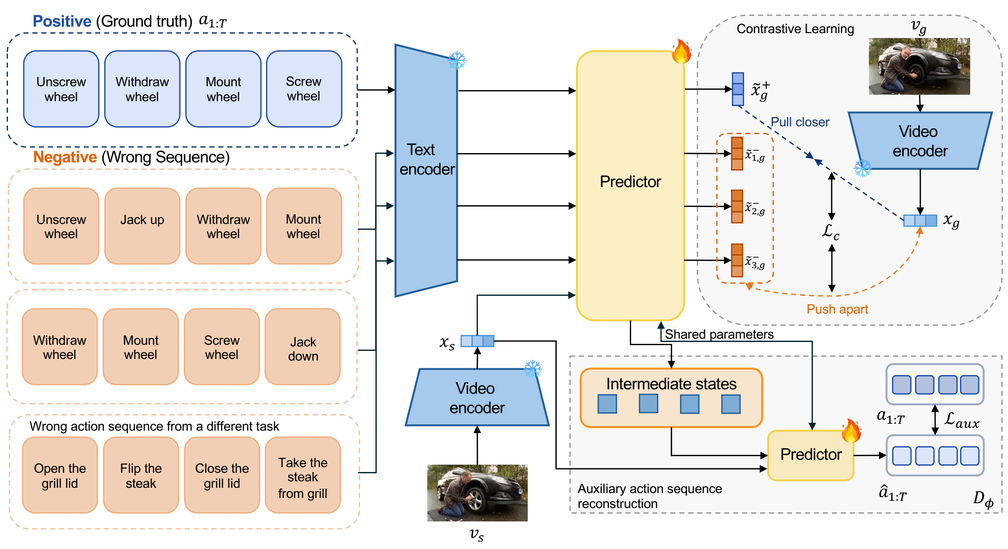


Fig. 1. Training. The ground-truth sequence (**blue**) is pulled towards x_g; N wrong sequences (**orange**) are pushed apart.

- **Contrastive triplet loss** with per-negative margin τ_i: The predicted goal for the ground-truth plan, $\tilde{x}_g^+$, is pulled towards x_g, while those for wrong plans, $\tilde{x}_{g,i}^-$, are pushed away.

$$\mathcal{L}_{\mathrm{c}} = \frac{1}{N}\sum_{i=1}^{N} \max\Big(\big\|x_g - \tilde{x}_g^+\big\|_2 - \big\|x_g - \tilde{x}_{g,i}^-\big\|_2 + \tau_i,\, 0 \Big)$$

- **Mixed negatives:** a mix of *hard* negatives (same task, wrong order or composition) and *easy* negatives (other tasks) combined in a fixed ratio
- **Adaptive margins:** negatives are not equally wrong, so each margin scales with the fraction of a negative's actions $\mathcal{A}_i^-$ that are absent from the ground-truth set $\mathcal{A}^+$:

$$\tau_i = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \frac{\big|\mathcal{A}_i^- \setminus \mathcal{A}^+\big|}{\big|\mathcal{A}_i^-\big|}$$

  so a reordering needs only $\tau_{\min}$, an unrelated sequence $\tau_{\max}$
- **Auxiliary loss:** intermediate states are never observed, so the *same* predictor runs twice with shared weights: mask tokens first yield the intermediate latents $\{\tilde{x}_l\}_{l=1}^{\mathrm{T}-1}$, which are then fed back with the actions masked so a decoding head D_ϕ reconstructs them under a cross-entropy term $\mathcal{L}_{\mathrm{aux}}$, giving $\mathcal{L} = \mathcal{L}_{\mathrm{c}} + \lambda\,\mathcal{L}_{\mathrm{aux}}$. Reconstruction succeeds only if those latents stay informative

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008). This project is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC-3066/1 "The Adaptive Mind", Project No. 533717223, EXC-3057/1 "Reasonable Artificial Intelligence", Project No. 533677015). Mohamed Afham & Christoph Reich are supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. This work was also supported by the ERC Advanced Grant SIMULACRON, the Georg Nemetschek Institute project AI4TWINNING, and the DFG project 4D-YouTube CR 250/26-1.

Method: Task-Constrained Inference

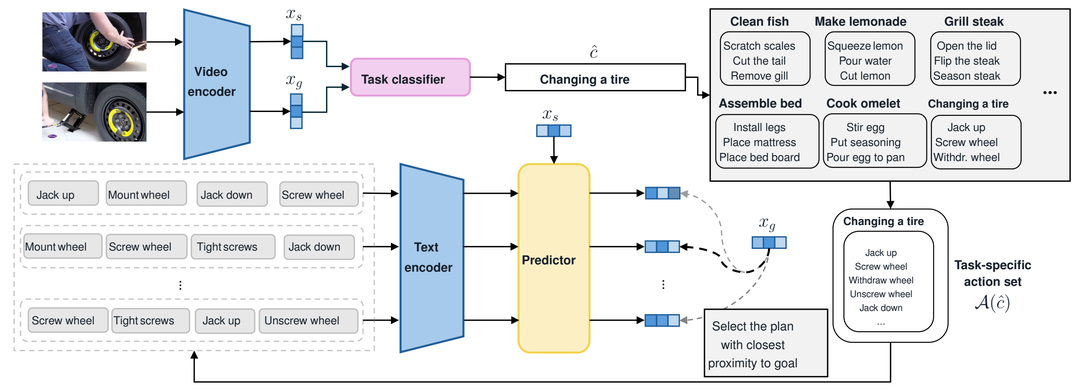


Fig. 2. Inference. A task classifier predicts the task $\hat{c}$ and restricts the search to its action set $\mathcal{A}(\hat{c})$; the candidate with the lowest energy is the predicted plan.

Planning is formulated as a discrete search: among the candidate plans that the task-specific action set $\mathcal{A}(\hat{c})$ allows, CEFITO returns the one whose predicted goal lies closest to the observed goal x_g.

$$\tilde{a}_{1:\mathrm{T}} \;=\; \underset{a_{1:\mathrm{T}} \in \mathcal{A}(\hat{c})}{\arg\min} \big\| P_\theta(x_s, a_{1:\mathrm{T}}) - x_g \big\|_2$$

Results

- **Benchmarks:** CrossTask [9] (2 750 videos, 18 tasks, 105 actions) and COIN [6] (11 827 videos, 180 tasks, 778 actions), under the unified protocol of [5]: five seeds, reporting SR, mAcc and mIoU
- **Setup:** frozen S3D visual features and a frozen CLIP text encoder; the predictor is a 4-layer transformer (6 heads, width 384, 8.5 M parameters), trained with 50 negatives per sample at a hard/easy ratio $r = 0.8$

Tab. 1. Comparison with the state of the art on CrossTask [9] (18 tasks) and COIN [6] (180 tasks) for horizons T = 3, 4 (in %, ↑; protocol of [5], mean and 90 % confidence interval over five seeds). SR: exact plan match. Best results highlighted in red, second best in orange.

	T = 3			T = 4		
Method	SR	mAcc	mIoU	SR	mAcc	mIoU
CrossTask [9]						
PDPP [7]	36.73 ±0.59	61.96 ±0.59	83.20 ±0.33	21.47 ±2.09	55.66 ±1.64	80.68 ±0.83
PlanLLM [8]	36.84 ±1.21	61.56 ±1.03	83.23 ±0.53	22.91 ±1.39	55.29 ±1.54	81.03 ±0.47
SCHEMA [4]	37.24 ±0.60	62.69 ±0.28	83.94 ±0.23	24.18 ±0.47	57.02 ±0.64	81.46 ±0.19
ViterbiPlanNet [5]	38.45 ±0.32	63.07 ±0.17	83.89 ±0.16	24.64 ±0.30	57.00 ±0.42	81.18 ±0.44
CEFITO (Ours)	39.62 ±0.24	64.12 ±0.31	84.29 ±0.21	24.76 ±0.43	57.53 ±0.37	81.58 ±0.25
COIN [6]						
PDPP [7]	22.37 ±0.57	44.60 ±0.16	83.00 ±0.42	15.21 ±0.34	41.01 ±0.32	81.64 ±0.48
PlanLLM [8]	33.44 ±0.15	51.05 ±0.46	84.66 ±0.41	23.19 ±0.32	45.70 ±0.33	83.44 ±0.39
SCHEMA [4]	32.89 ±0.61	50.84 ±0.47	83.98 ±0.67	22.33 ±0.92	45.21 ±1.05	82.93 ±0.25
ViterbiPlanNet [5]	33.99 ±0.23	50.87 ±0.17	83.88 ±0.31	23.92 ±0.29	45.63 ±0.55	82.56 ±0.44
CEFITO (Ours)	34.11 ±0.24	51.18 ±0.22	84.70 ±0.47	24.25 ±0.34	46.13 ±0.63	83.24 ±0.26

Tab. 2. Ablation. Component-wise analysis of CEFITO on CrossTask (in %, ↑)

	T = 3			T = 4		
Configuration	SR	mAcc	mIoU	SR	mAcc	mIoU
Baseline (L2 regression)	13.65	47.82	69.53	9.24	35.44	67.76
+ triplet contrastive loss	37.58	63.21	83.67	23.54	56.18	80.91
+ adaptive margin	38.74	63.79	84.02	24.23	57.01	81.22
+ auxiliary loss (CEFITO)	39.62	64.12	84.29	24.76	57.53	81.58

Task **Grill steak**



Start → Goal

Ground truth

Flip steak, Take steak from grill, Cut steak

Top-5 candidate plans	**Energy ↓**
Flip steak, Take steak from grill, Cut steak	**1.36**
Take steak from grill, Cut steak, Cut steak	1.43
Check temperature, Take steak from grill, Cut steak	1.45
Flip steak, Flip steak, Take steak from grill	1.48
Flip steak, Flip steak, Cut steak	1.50

Fig. 3. Qualitative result on CrossTask: candidate plans ranked by energy (L_2 distance to x_g); the top-ranked plan matches the ground truth.

Limitations: the search cost grows exponentially with the horizon T, so long-horizon inference is slow; accuracy drops on NIV, a benchmark two orders of magnitude smaller, where the contrastive objective has too few videos per task.

Conclusion

- **Contrastive learning** shapes an energy field over state transitions, conditioned on the action sequence: predicted transitions following the ground-truth actions land closer to the true goal, while incorrect sequences are pushed away
- **Task-constrained search** at inference time returns the lowest-energy plan, restricted to the actions of the predicted task
- **State-of-the-art** results on CrossTask and COIN

References

1. M. Botvinick and M. Toussaint, "Planning as Inference," in *Trends in Cognitive Sciences*, 2019.
2. C.-Y. Chang *et al.*, "Procedure Planning in Instructional Videos," in *ECCV*, 2020.
3. K. R. Y. Nagasinghe *et al.*, "Why Not Use Your Textbook? Knowledge-Enhanced Procedure Planning of Instructional Videos," in *CVPR*, 2024.
4. Y. Niu *et al.*, "SCHEMA: State Changes Matter for Procedure Planning in Instructional Videos," in *ICLR*, 2024.
5. L. Seminara *et al.*, "ViterbiPlanNet: Injecting Procedural Knowledge via Differentiable Viterbi for Planning in Instructional Videos," in *CVPR*, 2026.
6. Y. Tang *et al.*, "COIN: A Large-Scale Dataset for Comprehensive Instructional Video Analysis," in *CVPR*, 2019.
7. H. Wang *et al.*, "PDPP: Projected Diffusion for Procedure Planning in Instructional Videos," in *CVPR*, 2023.
8. D. Yang *et al.*, "PlanLLM: Video Procedure Planning with Refinable Large Language Models," in *AAAI*, 2025.
9. D. Zhukov *et al.*, "Cross-Task Weakly Supervised Learning from Instructional Videos," in *CVPR*, 2019.

afham.aflal@visinf.tu-darmstadt.de