University of Science and Technology of China

Target-Driven Policy Optimization for Sequential Counterfactual Outcome Control

GIFT: Goal-conditioned Intervention via Factual-Targeted Training

Problem

Target-reaching control from observational data.

Insight

Identifiable and stable goal-conditioned RL.

Method

HER + clipped reward rescaling for SAC.

Authors: Xin Wang, Xiangyu Zhang, Shengfei Lyu, Huanhuan Chen

Presenter: Xin Wang

University of Science and Technology of China; Nanyang Technological University

Talk Structure

problem, theory, method, evidence.

01 Background
Why open-loop plans are brittle

02 Formulation
SCTA as a goal-conditioned MDP

03 Theory
Identifiability, contraction, bias

04 Method
HER + reward-rescaled SAC

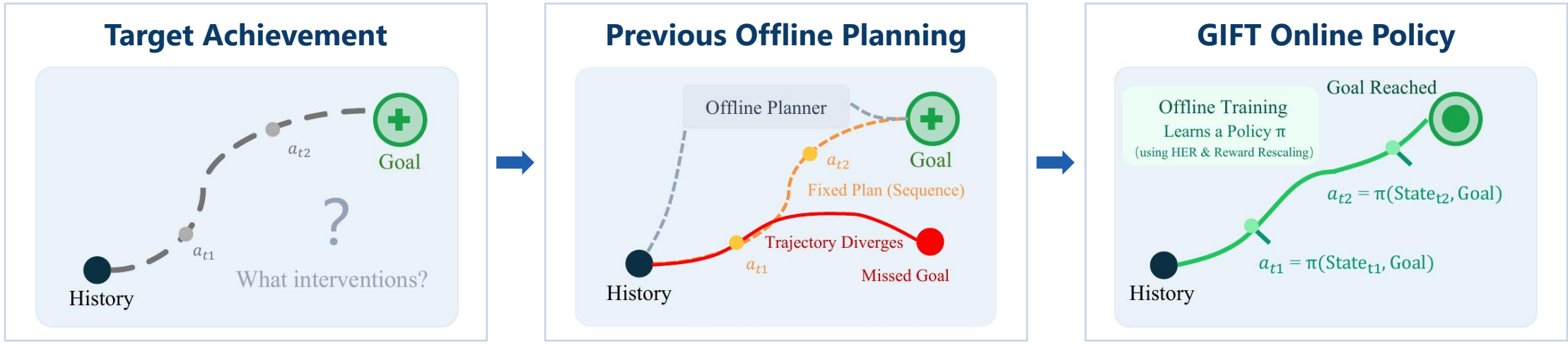
05 Experiments
Tumor and MIMIC-III synthetic

06 Results
Accuracy, generalization, efficiency

07 Conclusion
Take-home message

Motivation: From Open-Loop Plans to Reactive Policies

Motivation schematic: fixed plans fail when trajectories deviate.



Three obstacles in observational cohorts

Confounding

Treatment choices depend on patient history.

Distribution Shift

The learned policy may leave data support.

Sparse Rewards

Target hits are rare in logged trajectories.

Central shift: from counterfactual prediction to adaptive target-reaching policies.

SCTA as a Goal-Conditioned MDP

The policy is learned from longitudinal observational data.

Data

Covariates, actions, outcomes, static features.

$$\Psi_t = (\bar{H}_t, Y_{target})$$

State = history + target.

$$\mathcal{T} = \{ y : \|y - Y_{target}\| \leq \delta \}$$

Target tolerance region.

Reward

Step penalty until the first target hit.

$$V^{\pi}(\psi_t) = E_{\pi}[\sum_{k=0}^{\tau_{max}-1} \gamma^k r_{t+k}]$$

Objective: maximize discounted target-achievement return.

Assumptions

Consistency, ignorability, positivity, goal independence.

Policy surface
Input: history + target
Output: next continuous intervention

Stable Offline Learning via Clipped Rewards

Clipped importance weights stabilize offline Bellman learning.

Identifiability

Bellman targets are estimable from observational transitions.

Reward Rescaling

Bounded ratios move rewards toward the actor policy.

Guarantee

Contraction plus explicit clipping-bias bound.

$$\rho(\psi, a) = \frac{\pi_\theta(a|\psi)}{\pi_b(a|\psi)}, \quad \bar{\rho} = \mathrm{clip}(\rho, \varepsilon_1, \varepsilon_2)$$

$$\tilde{r}(\psi, a) = \bar{\rho}(\psi, a)\, r(\psi, a)$$

$$\|T_{\tilde{\pi}} Q_1 - T_{\tilde{\pi}} Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$$

$$\|Q^{\pi_\theta} - Q^{\tilde{\pi}}\|_\infty \leq \frac{1}{1-\gamma} \|(T^{\pi_\theta} - T_{\tilde{\pi}}) Q^{\tilde{\pi}}\|_\infty$$

Takeaway: lower variance, stable learning, controlled bias.

GIFT: Factual-Targeted Training

GIFT couples target relabeling, representation learning, and offline SAC.

1 Offline Data
Histories
actions
outcomes

2 HER
Relabel achieved
future outcomes

3 Encoder
LSTM history
+ MLP goal

4 SAC
Train with
clipped ratios

5 Rollout
Update history
choose action

$$\pi_\theta(a_t \mid S_t, G_t)$$

closed-loop execution

Why it fits SCTA
HER: sparse targets
Clipping: offline shift
SAC: continuous actions

Inference advantage: direct policy rollout replaces costly test-time sequence optimization.

Setup and Evaluation Protocol

Evaluation tests target-reaching accuracy, generalization, and efficiency.

Datasets

Tumor simulator
MIMIC-III synthetic

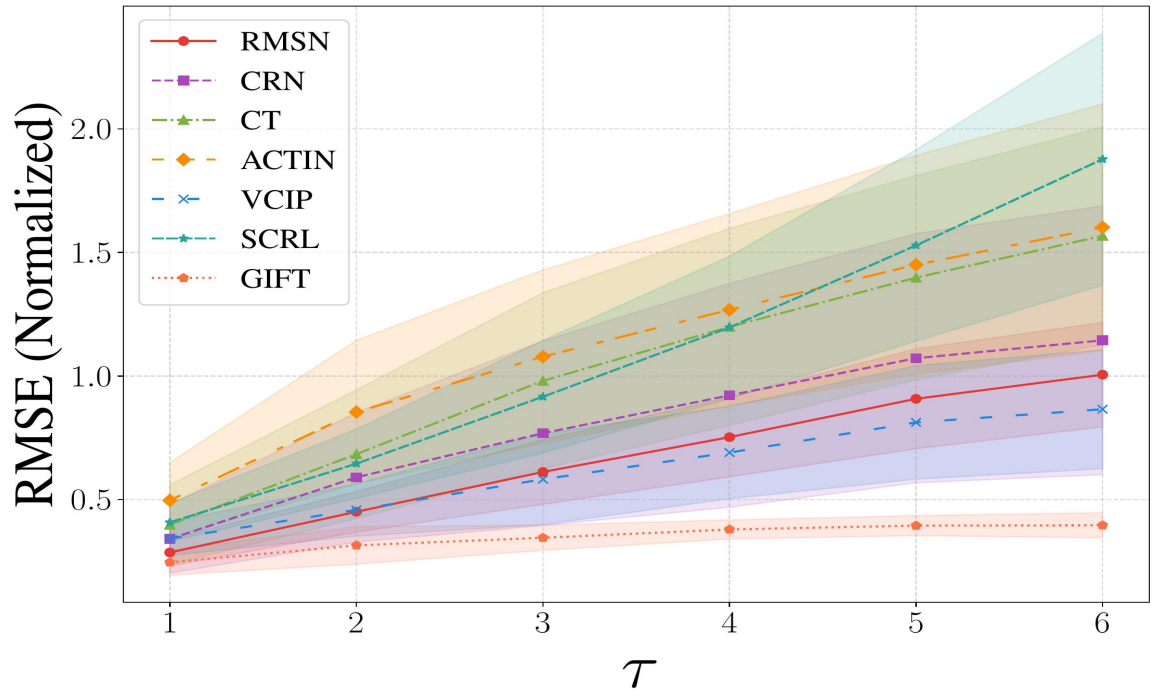
Baselines

RMSN, CRN, CT,
ACTIN, VCIP, SCRL

$$\mathrm{RMSE} = \sqrt{\frac{\sum \|\mathrm{Y}_{\mathrm{term}} - \mathrm{Y}_{\mathrm{target}}\|_2^2}{\mathrm{N}}}$$

Terminal RMSE after rollout.
Lower is better.

Paper line chart: Tumor dataset under stronger confounding



Core question: Can an online policy hit targets more accurately than offline planning?

Main Results and Analysis

GIFT lowers terminal RMSE and keeps inference efficient.

Representative Test Performance

Setting	GIFT	Best Baseline	Gain
MIMIC, τ=6	0.39	0.65	40%
Tumor κ=4, τ=6	0.46	1.10	58%
Tumor κ=4, τ=3	0.38	0.69	45%

Ablation Summary

Variant	MIMIC	Tumor κ=4
Full GIFT	0.39	0.40
w/o HER	0.95	1.53
CQL	0.49	1.04

Effectiveness

Consistently lower RMSE across horizons.

Generalization

Targets from unseen strategies remain easier to reach.

Efficiency

Policy rollout avoids iterative planning.

Main observation: HER supplies target hits; clipped reward rescaling stabilizes actor-critic learning.

Conclusion

GIFT turns sequential counterfactual target achievement into adaptive, goal-conditioned policy learning.

Formulation

Goal-conditioned MDP for target control.

Algorithm

HER + reward-rescaled offline SAC.

Theory

Contraction and bounded clipping bias.

Evidence

Better accuracy and faster inference.

Take-home message

Goal-conditioned offline RL can replace brittle open-loop intervention planning when identifiability, HER, and variance control are handled together.

Limitations

Causal assumptions matter.
Clipping needs careful tuning.

Future Work

Certified policy evaluation.
Prospective clinical validation.

Authors: Xin Wang, Xiangyu Zhang, Shengfei Lyu, Huanhuan Chen | Presenter: Xin Wang