Counterfactual Reasoning: Theory and Applications



Pusan National University
Computer Vision & Signal Processing Lab, Dept. of Electronic Engineering



Presented By
김소형

CVSP Lab
Computer Vision & Signal Processing

Index

- 논문 선정 이유
- 논문 소개
  - Reasoning with cause and effect [IJCAI 1999]
  - Should I Have Expressed a Different Intent? Counterfactual Generation for LLM-Based Autonomous Control [ICML 2026]
- Conclusion

Motivation for Paper Selection

- Reasoning with cause and effect [IJCAI 1999]
  - Counterfactual의 개념을 보다 명확히 이해하고자 선정함
    - 우리 알고리즘에 Counterfactual의 개념을 적용하기 위함



- Should I Have Expressed a Different Intent? Counterfactual Generation for LLM-Based Autonomous Control [ICML 2026]
  - Counterfactual의 근본 개념을 활용한 사례를 참고하고자 선정함
    - Counterfactual의 근본 개념을 딥러닝에 적용하는 논리와 실험적 근거를 참고하고자 함

Who is Judea Pearl?

- 참고



Judea Pearl

팔로우

Professor of Computer Science, University of California, Los Angeles
cs.ucla.edu의 이메일 확인됨 - 홈페이지

causality artificial intelligence structural equations counterfactuals

인용	전체	2021년 이후
서지정보	176962	67258
h-index	130	82
i10-index	398	244

모두 보기

제목	인용	연도
Probabilistic reasoning in intelligent systems: Networks of plausible inference J Pearl Synthese-Dordrecht 104 (1), 161	36048 *	1995
Causality J Pearl Cambridge university press	34850	2009
The book of why J Pearl, D Mackenzie Allen Lane 14 (31), 47-54	6368	2018
Heuristics: intelligent search strategies for computer problem solving J Pearl Addison-Wesley Pub. Co., Inc., Reading, MA	5574	1983

Turing Award

ACM A. M. Turing Award

튜링상

2011년	주데아 펄	인공지능, 확률적 알고리즘과 원인 추론

CVSP Lab
Computer Vision & Signal Processing

Are Counterfactuals Effective?

- 참고
  - 최신 연구에서 Counterfactual 키워드 논문이 늘어나는 추세임

ICML 2026	Poster	25편
ICLR 2026	Poster 18 + Oral 1	19편
CVPR 2026	Poster 13 + Highlight 2	15편
ICCV 2025	Poster 5 + Highlight 1	6편
총합		65편

Reasoning with Cause and Effect [IJCAI 1999]

CVSP Lab
Computer Vision & Signal Processing

Introduction

Pearl, Judea. "Reasoning with cause and effect." *AI Magazine* 23.1 (2002): 95-95.

- Abstract
  - 이 논문의 중심 주제는 인과성(causality)을, 어떤 영역에서 불변하는 관계(invariant relationships)를 식별하여 행동(actions)의 효과를 예측하는 것을 용이하게 하기 위해 고안된 계산적 체계(computational scheme)로 해석하는 것이다.

CVSP Lab
Computer Vision & Signal Processing

Introduction

Pearl, Judea. "Reasoning with cause and effect." *AI Magazine* 23.1 (2002): 95-95.

- 연구배경
  - 기존 연구자
    - Counterfactual를 평가할 때 Similarity에 의존하는 방법
      - "모든 world는 자기자신과 가장 가깝다" – Semantic
        - ✓ex) 만약 A였다면, B였을 것이다 → 실제세계 w에서 참

```
Actual world w
      ↓
A가 참인 가능한 세계들을 찾음
      ↓
그중 w와 가장 비슷한 세계들을 선택
      ↓
그 모든 세계에서 B가 참인가?
      ↓
Yes → "If A, then B" 참
```

  - Judea Pearl
    - Counterfactual을 world간 Similarity가 아닌 causal structure의 수정으로 정의

CVSP Lab
Computer Vision & Signal Processing

Method

Pearl, Judea. "Reasoning with cause and effect." *AI Magazine* 23.1 (2002): 95-95.

- Structural Causal Model (**SCM**)
  - $M = < U, V, F >$
    - U: Exogenous variables
      - 모델 밖의 요인에 의해 결정되는 배경조건
      - 개체, 상황, 환경의 구체적인 상태
    - V: Endogenous variables
      - 모델 안에서 결정되는 변수
      - 다른 변수와 U의 영향을 받아 값이 정해짐
    - F: Structural functions
      - $F = \{f_1, f_2, \dots, f_n\}$

CVSP Lab
Computer Vision & Signal Processing

Method

Pearl, Judea. "Reasoning with cause and effect." *AI Magazine* 23.1 (2002): 95-95.

- Abduction-Action-Prediction
  - 주어진 모델
    - $< M, P(u) >$
  - Abduction
    - $P(u) \rightarrow P(u|e)$
      - Factual evidence로 어떤 배경상황이 실제 episode를 만들었는지 추론
  - Action
    - $M \rightarrow M_A$
      - Counterfactual antecedent $X = x$를 intervention으로 적용
  - Prediction
    - $< M_A, P(u|e) >$

CVSP Lab
Computer Vision & Signal Processing

Should I Have Expressed a Different Intent? Counterfactual Generation for LLM-Based Autonomous Control

ICML 2026

Introduction

Farzaneh, Amirmohammad, Salvatore D'Oro, and Osvaldo Simeone. "Should I Have Expressed a Different Intent? Counterfactual Generation for LLM-Based Autonomous Control." *arXiv preprint arXiv:2601.20090* (2026).

- Motivation
  - LLM 기반 에이전트는 사용자의 high-level intent를 환경에서 수행할 수 있는 계획과 행동으로 변환할 수 있음
    - 결과를 관찰한 뒤, 사용자는 다음과 같은 의문을 가질 수 있음
      - " 내가 의도를 다르게 표현했다면 어떻게 되었을까?"
  - 다른 intent를 사용했을 때의 counterfactual outcome을 신뢰성 있게 예측하는 방법 필요

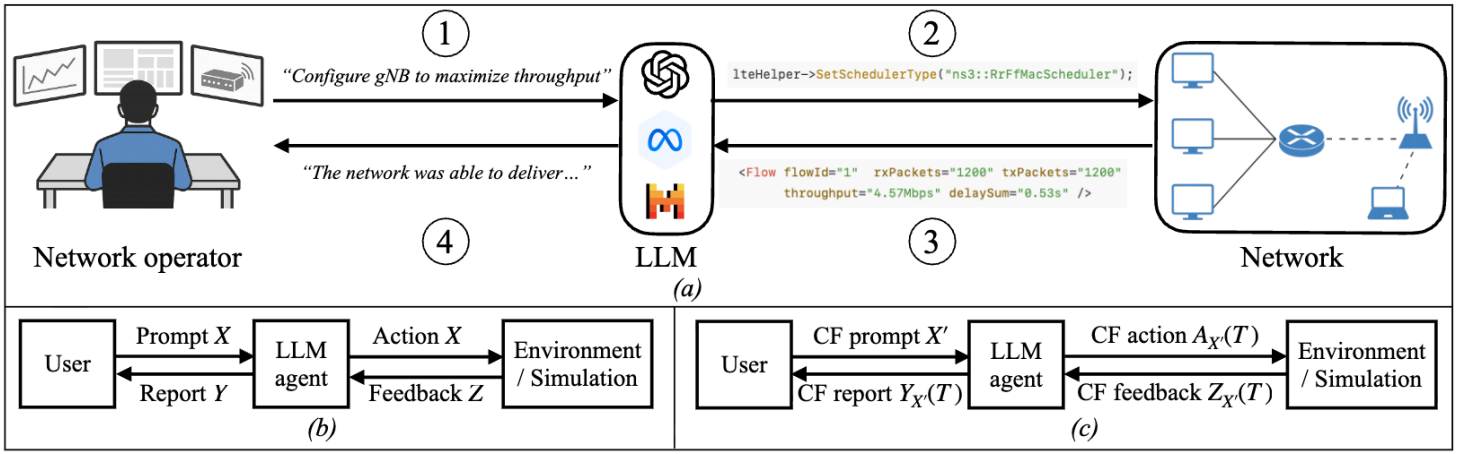
Introduction

Farzaneh, Amirmohammad, Salvatore D'Oro, and Osvaldo Simeone. "Should I Have Expressed a Different Intent? Counterfactual Generation for LLM-Based Autonomous Control." *arXiv preprint arXiv:2601.20090* (2026).

- Contribution
  - Agent–Environment Counterfactual Generation
    - 전체 agent–environment interaction을 **SCM(Structural Causal Model)**으로 모델링
  - Experimental Results
    - Naive re-execution 대비 더 정확하고 안정적인 counterfactual generation

Introduction

- Problem Definition
  - 용어정의
    - 에피소드 $T = (X, A, Z, Y)$
      - X: 사용자 프롬프트
      - A: 에이전트가 생성하여 환경에서 실행되는 일련의 행동
      - Z: 에이전트의 행동에 대한 환경의 반응을 설명하는 환경 피드백
      - Y: 피드백 Z의 함수로서 사용자에게 반환되는 텍스트 보고서
    - Counterfactual prompt $X' \in E(X)$
      - $E(X)$: 주어진 factual prompt X에 대한 counterfactual prompt 집합

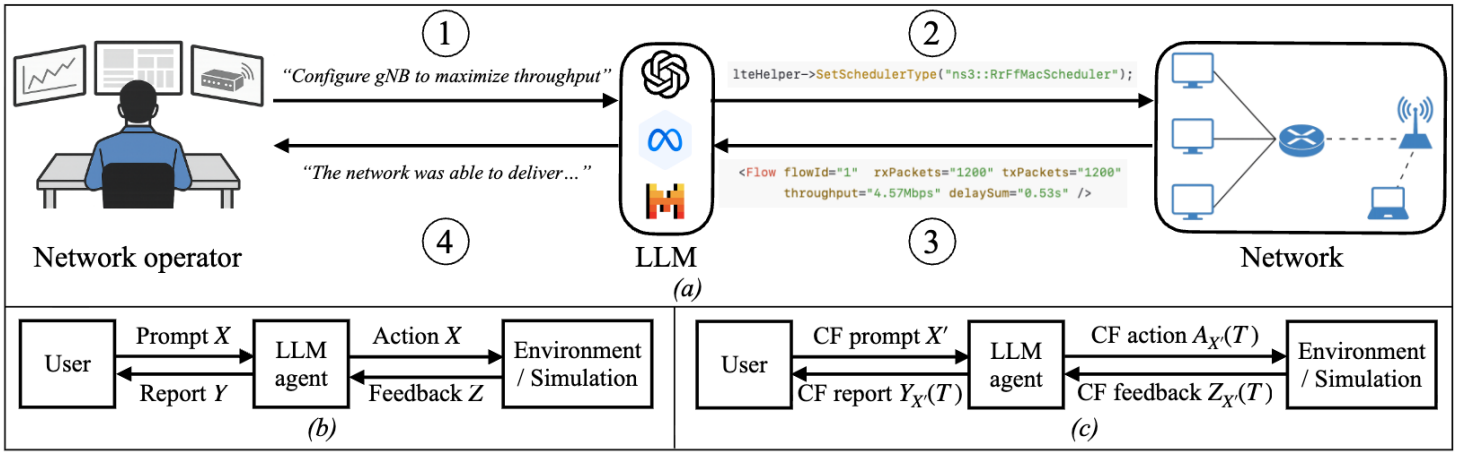


CVSP Lab
Computer Vision & Signal Processing

Introduction

- Problem Definition
  - 목표
    - 입력 프롬프트가 였다고 가정했을 때, 동일한 조건하에서 사용자가 받았을 counterfactual report $Y_{X'}(T)$를 추정하는 것

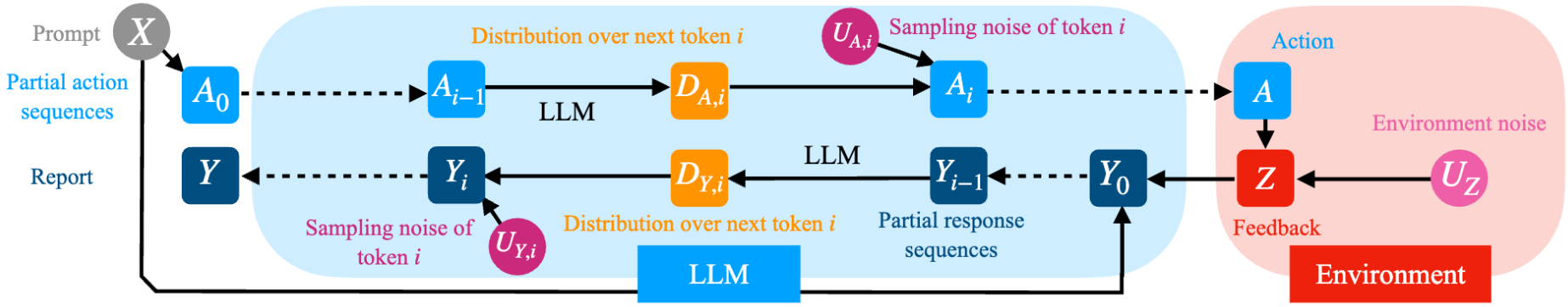
Factual intent (X)	Counterfactual intent (X')
Context: 5 UEs with **time-varying** SNRs. Each UE has expected offered load 1.20 Mbps. Goal: Configure gNB to **maximize aggregate throughput** while promoting fairness across UEs. Targets: Per-UE throughput around **1.08 Mbps** on average; average per-UE latency ≤ **18 ms.** Summarize the KPIs over a period of 10 seconds.	Context: 5 UEs with **static** SNRs. Each UE has expected offered load **1.20 Mbps**. Goal: Configure gNB to **maximize fairness** and latency stability. Targets: Per-UE throughput around **0.84 Mbps** on average; average per-UE latency ≤ 10 ms. Summarize the KPIs over a period of 10 seconds.
Factual report (Y)	**True counterfactual report ($Y_{X'}$)**
Chosen scheduler: PF. Over 10 s, the system achieves an average throughput of 1.07 Mbps per UE, with total throughput around 5.3–5.5 Mbps for most of the run. The average end-to-end latency remains between 15–18 ms. A brief dip occurs at $t \approx 4.4$ s, where total throughput drops to about 5.0 Mbps before recovering quickly.	Chosen scheduler: RR. The system delivers stable throughput at approximately 0.84 Mbps per UE (total ≈4.2 Mbps), with very small spread across users. The average latency stays tightly within 8–10 ms. A short dip is observed near $t \approx 4.5$s, where total throughput decreases to about 3.9 Mbps, followed by a smooth return to the previous level.
Interventional generation (IG)-based report ($\hat{Y}_{X'}$)	**Counterfactual generation (CG)-based report ($\hat{Y}_{X'}(T)$)**
Chosen scheduler: RR. In this run, total throughput varies between 3.4–5.0 Mbps, with per-UE rates ranging from 0.65–1.05 Mbps. Latency fluctuates between 7–14 ms, with several intervals above the 10 ms target. A	Chosen scheduler: RR. Per-UE throughput remains close to 0.84 Mbps (total ≈4.2 Mbps) with uniform allocation. Average latency stays low (8–9 ms). A mild dip appears near $t \approx 4.4$ s before fast recovery.



CVSP Lab
Computer Vision & Signal Processing

Method

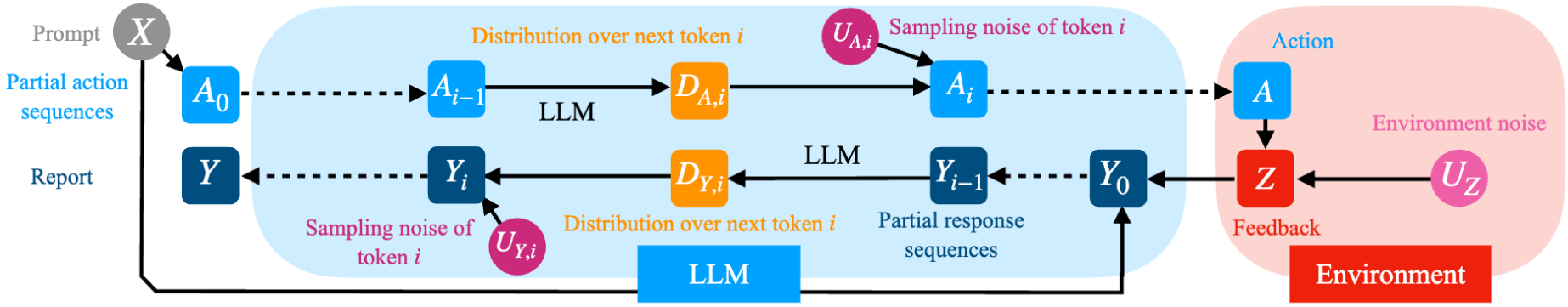
- Modeling the Agent–Environment System as a Structural Causal Model
  - Action generator
    - Next-token probability distribution
      - LLM의 다음 token에 대한 probability distributioion $P_{A,i} = [P_{A,i,1}, \dots, P_{A,i,V}]^T$
        - ✓A: LLM의 Action, i: action sequence, V : LLM vocabulary의 크기
    - Next-token sampling via Gumbel-Max
      - Exogenous noise: $U_{A,i} = [U_{A,i,1}, \dots, U_{A,i,V}] \sim Gumble(0,1)$
        - ✓$U_{A,i}$: 크기가 V x 1 벡터인 Random noise set
      - Next token $A_i = arg \max_{v \in \{1,\dots,V\}} \{\log P_{A,i,v} + U_{A,i,v}\}$



CVSP Lab
Computer Vision & Signal Processing

Method

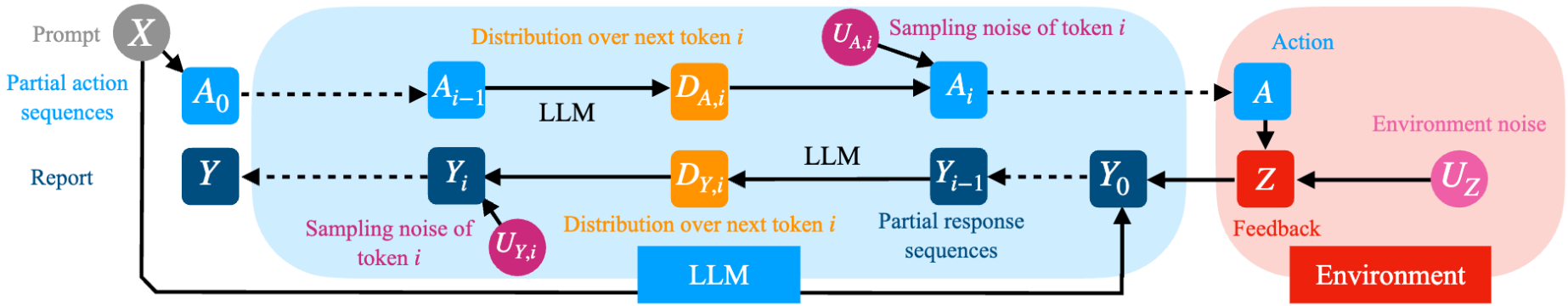
- Modeling the Agent–Environment System as a Structural Causal Model
  - Environment model
    - $Z = f_Z(A, U_Z)$ → Action의 효과와 외부요인을 분리하여 counterfactual 적용
      - A : LLM이 생성한 실행 설정, U_Z: Action으로 설명되지 않는 환경 조건
      - Z: output, f_Z: 환경의 작동 메커니즘
  - Report generator
    - Action generator에서 (X, A, Z)를 조건으로 한다는 것이 다름
    - $U_y = (U_{Y,1}, U_{Y,2}, \ldots)$
      - U_y: report Y를 생성할 때 사용된 모든 token-step의 sampling noise set



CVSP Lab Computer Vision & Signal Processing

Method

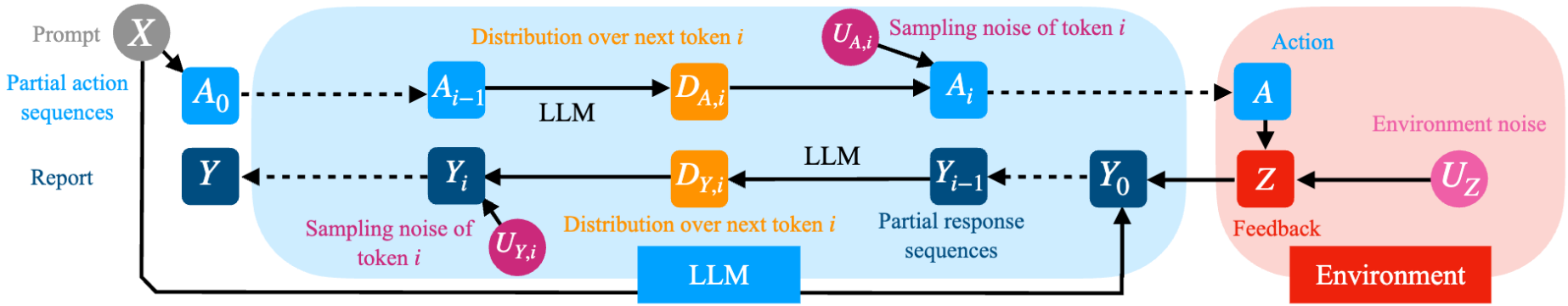
- Counterfactual Generation
  - Prompt Editing
    - Factual prompt X가 주어지면, 원래 intent를 수정하여 counterfactual prompt X' 를 지정
  - Abduction
    - 관측된 factual episode T로부터, 시스템이 exogenous noise variable U_A, U_Y, U_Z를 추론
      - U_A, U_Y는 앞서 Gumble Max sampling으로 얻어진 결과이기에 known
      - U_Z는 latent variable
        - ✓ 관측된 trace인 $T = (X, A, Z, Y)$로부터 추론하여야 함 → $\hat{U}_Z$



CVSP Lab
Computer Vision & Signal Processing

Method

- Counterfactual Generation
  - Action Counterfactual Generation
    - counterfactual prompt X'와 U_A를 사용해 다시 실행하여 $A_{X'}(T)$를 생성
  - Environment Counterfactual Simulation
    - $A_{X'}(T)$와 추론된 $\hat{U}_Z$로 $\hat{Z}_{X'}(T) = f_z(A_{X'}(T), \hat{U}_Z)$를 생성
  - Report Counterfactual Generation
    - Report를 생성하는 LLM에 $(X', A_{X'}(T), \hat{Z}_{X'}(T))$와 U_Y를 사용하여 $\hat{Y}_{X'}(T)$ 생성



CVSP Lab
Computer Vision & Signal Processing

Experiment

- Fidelity in reconstructing the counterfactual KPI time series(best in **bold**)
  - IG(Interventional Generation)
    - Counterfactual prompt X'를 실제 시스템에 바로 입력
      - $\hat{Z}_{X'}^{IG} = f_{real}(A_{X'}, U'_Z)$
  - SIG(Simulated interventional Generation)
    - IG와 동일하게 factual을 무시하고 새로운 U_Z를 prior에서 뽑음
      - $\hat{Z}_{X'}^{SIG} = f_Z(A_{X'}, U'_Z)$
  - CG(Counterfactual Generation)
    - Factual observatioin A, Z를 이용하여 환경조건을 추론
      - $\hat{Z}_{X'}^{CG} = f_Z(A_{X'}(T), \hat{U}_Z)$

MAE: GT Z_t와 비교
Cross-corr.: 증가, 감소 패턴이 얼마나 비슷한지
Crossing-level: 중요한 임계값을 초과한 시간비율을 얼마나 정확히 맞췄는가

	CG / IG / SIG		
Metric	**MAE**	**Cross-corr. peak**	**Crossing-level error**
Throughput	**0.15** / 0.28 / 0.33	**0.93** / 0.78 / 0.70	**0.03** / 0.14 / 0.18
Delay	**0.35** / 0.52 / 0.60	**0.89** / 0.71 / 0.62	**0.05** / 0.20 / 0.26

CVSP Lab
Computer Vision & Signal Processing

Thank you

https://annon-repo-for-science.github.io/CounterFactual/