

Causal Inference with Video Features as Treatments

Kosuke Imai

Harvard University

Techstars Tokyo

October 19, 2026

Joint work with Kentaro Nakamura, Adam Breuer, Bryce Dietrich, and Mike Crespín

AI and Causality

- Causal inference
 - different from prediction
 - action vs. observation
 - for decision making, we need causality not prediction
 - simply predicting observed data does not let you identify causal relations
- AI and causality
 - How to leverage AI for causal inference (today)
 - How to debias AI predictions for decision-making (on-going)
 - How to incorporate causality into AI models (future)

GenAI-Powered Inference (GPI) (Imai and Nakamura, *PNAS* 2026, *JASA* in-press)

- Causal inference framework for unstructured data (text, image, video, etc.)
 - ① leverage a GenAI model to (re)generate the unstructured data
 - ② extract its **internal representations**
 - ③ estimate feature-specific causal effects using extracted representations
- Advantages
 - ① representations contain *all* the information of generated objects in a **low-dimensional** format
 - ② use **neural network** to avoid functional-form assumptions
 - ③ asymptotic **statistical guarantees** with good performance under moderate sample sizes
- Applications
 - ① text features as treatment
 - ② image features as treatment
 - ③ **video features as treatment** (arXiv preprint)

Motivation: Video-as-treatment

- Video is a dominant medium of modern communication
- How do we conduct **causal inference** with video?
- Standard approach:
 - ① classify videos into treatment and control groups based on features of interest
 - ② randomly assign viewers to videos
 - ③ compare viewers' reactions to treated and control videos
- Challenge: treatment features are **confounded** with many other visual and auditory features
- How can we identify the causal effects of specific video features?
- How can we estimate their dynamic effects on viewers' reactions over time?
- To our knowledge, no causal inference method is available for video features

Empirical Application: Presidential Campaign Advertisement

- How does the candidate's appearance affect a voter's real-time evaluation?
- Data: 2020 campaign videos from the Wesleyan Media Project
 - 849 videos (10–120 seconds; median: 30 seconds)
 - Joe Biden (389 videos) / Donald Trump (128 videos)
- **Treatment feature:** candidate appearance
 - online survey conducted on Prolific
 - asked respondents to press the "R" key when candidates appeared
 - each video coded by up to five coders
 - videos divided into 5-second segments
 - a segment is "treated" if a majority of coders identify the candidate

Watch for this candidate on the screen:



BIDEN, JOE

Please click the R key each time you see this candidate appear on the screen.



TAP TO ENABLE AUDIO

is not playing. If the video is paused, click on the play button.

Real-time Evaluation of Campaign Advertisement

- separate online survey conducted on Prolific ($N = 1,108$)
- randomly assign videos to respondents
- respondents continuously evaluated the candidate using a feeling thermometer (0–100)
- dial positions were recorded throughout the video
- each respondent rated on average 9.81 advertisements

Rate the candidate who sponsored this ad:



Donald Trump

Use the slider below to rate how you feel about Donald Trump as you watch the ad.

Move it left for colder feelings and right for warmer feelings.



At this point in the ad, how would you rate the political candidate who sponsored the ad?

50 – Neither Cold Nor Warm



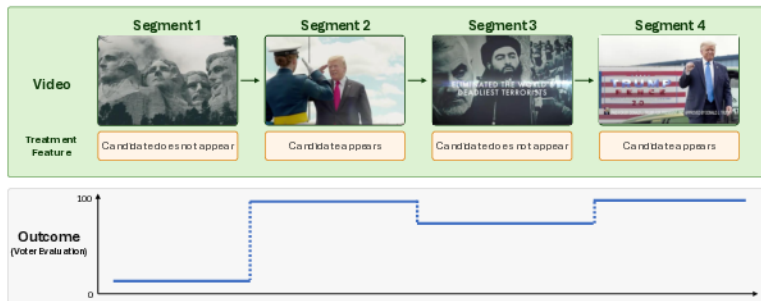
Video playing. Use the slider to rate the candidate as you watch.

Methodological Challenges

- ① High-dimensionality of videos
 - even a short video contains many audiovisual features
 - pre-trained embedding is less developed for video data and its quality is unknown
- ② Confounding by other video features
 - candidates do not appear in ads randomly
 - e.g., candidates are more likely to appear in ads with positive emotions
 - random assignment of videos is not enough
- ③ Dynamic real-time outcome
 - viewers' current evaluation might depend on any of the previous frames
 - unknown carryover effects

Can we estimate the causal effects of video features on real-time responses?

Setup: Segmenting Videos



- $\mathbf{X}_i$: videos assigned to individual i
- $\mathbf{X}_{is}$: s -th segment of video assigned to individual i
- S_i : length of video segments assigned to individual i (assumed to be fixed for now)
- Y_{is} : outcome of respondent i at segment s (e.g., voter evaluation)
- $Y_{is}(\mathbf{x}_1, \dots, \mathbf{x}_s)$: potential outcome for individual i at segment s
- W_{is} : treatment feature of videos assigned to individual i at segment s

Assumptions

- ① **Consistency:** $Y_{is} = Y_{is}(\mathbf{X}_{i1}, \dots, \mathbf{X}_{is})$
- ② **Random assignment of videos:** $Y_{is}(\mathbf{x}_1, \dots, \mathbf{x}_s) \perp\!\!\!\perp \mathbf{X}_i$
- ③ **Treatment Feature:** known, objective feature in $\mathbf{X}_{is}$

$$W_{is} = g_W(\mathbf{X}_{is}) \in \{0, 1\}$$

- ④ **Confounding Features:** unknown, objective features in $\mathbf{X}_{is}$ for the outcome $Y_{is'}$

$$\mathbf{U}_{is}^{(s')} = \mathbf{g}_{\mathbf{U}_s}^{(s')}(\mathbf{X}_{is})$$

for $s \leq s'$ where $\dim(\mathbf{U}_{is}^{(s')}) \ll \dim(\mathbf{X}_{is})$

5 Sequential Separability:

$$Y_{is}(\mathbf{x}_1, \dots, \mathbf{x}_s) = Y_{is}(g_W(\mathbf{x}_1), \mathbf{g}_{U_1}^{(s)}(\mathbf{x}_1), \dots, g_W(\mathbf{x}_s), \mathbf{g}_{U_s}^{(s)}(\mathbf{x}_s))$$

$$\mathbf{U}_{is'}^{(s)}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{s'-1}) = \mathbf{U}_{is'}^{(s)}(g_W(\mathbf{x}_1), \mathbf{g}_{U_1}^{(s)}(\mathbf{x}_1), \dots, g_W(\mathbf{x}_{s'-1}), \mathbf{g}_{U_{s'-1}}^{(s)}(\mathbf{x}_{s'-1}))$$

- we can intervene on the treatment feature while holding the confounding features constant
- implies independence of support (i.e., positivity given confounding features)

$$\mathcal{S}_{WU,s}^{(s')} = \mathcal{W}_s \times \mathcal{U}_s^{(s')}$$

$$\text{where } \mathcal{S}_{WU,s}^{(s')} := \left\{ \left(g_W(\mathbf{x}), \mathbf{g}_{U_s}^{(s')}(\mathbf{x}) \right) : \mathbf{x} \in \mathcal{X}_s \right\}$$

6 Factorized deterministic decoding for a GenAI video model

$$\mathbb{P}(\mathbf{X}_i \mid \mathbf{h}_\gamma(\mathbf{R}_i)) = \prod_{s=1}^{S_i} \mathbb{P}(\mathbf{X}_{is} \mid \mathbf{h}_\gamma(\mathbf{R}_{is})) \quad \text{is degenerate}$$

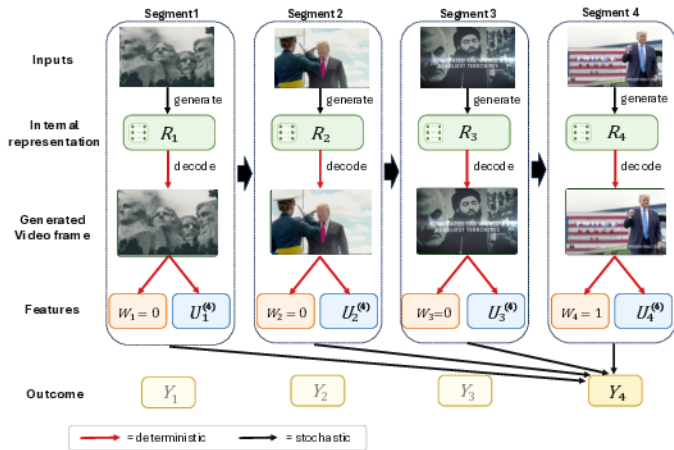
$$\mathbb{P}(\mathbf{R}_i \mid \mathbf{P}_i)$$

where $\mathbf{R}_{is}$ is the internal representation of $\mathbf{X}_{is}$ and $\mathbf{P}_i$ is the input to the diffusion model

Reproducing Video Data from Generative AI (NVIDIA Cosmos Tokenizer)



Summary of Assumptions



$$Y_{is} \perp\!\!\!\perp \{R_{i1}, \dots, R_{is}\} \mid W_{i1}, \dots, W_{is}, U_{i1}^{(s)}, \dots, U_{is}^{(s)}, S_i \geq s$$

$$U_{i,s'+1}^{(s)} \perp\!\!\!\perp \{R_{i1}, \dots, R_{is'}\} \mid W_{i1}, \dots, W_{is'}, U_{i1}^{(s)}, \dots, U_{is'}^{(s)}, S_i \geq s \text{ for } s' < s$$

Stochastic Intervention

- High-dimensional data $\rightsquigarrow$ violation of overlap assumption
- Want our counterfactuals to be close to the observed data
- Stochastic intervention: for each segment s ,

$$q_s(\delta_s; \bar{\mathbf{w}}_{s-1}) := \frac{\delta_s p_s(\bar{\mathbf{w}}_{s-1})}{\delta_s p_s(\bar{\mathbf{w}}_{s-1}) + 1 - p_s(\bar{\mathbf{w}}_{s-1})}, \quad \delta_s \in (0, \infty)$$

where $p_s(\bar{\mathbf{w}}_{s-1}) = \mathbb{P}(W_{is} = 1 \mid \bar{\mathbf{W}}_{i,s-1} = \bar{\mathbf{w}}_{s-1}, S_i \geq s)$ is the treatment probability

- a greater value of δ_s (odds ratio of q_s vs. p_s) $\rightarrow$ higher assignment probability
- q_s intervenes only when possible, given the treatment history
 - When $p_s(\bar{\mathbf{w}}_{s-1}) = 1$, then intervention q_s always assigns 1
 - When $p_s(\bar{\mathbf{w}}_{s-1}) = 0$, then intervention q_s always assigns 0
- **Bounded overlap** assumption: There exists $c > 0$ such that

$$\pi_s(\bar{\mathbf{w}}_{s-1}, \bar{\mathbf{u}}_s) \geq c \cdot p_s(\bar{\mathbf{w}}_{s-1}) \quad \text{and} \quad 1 - \pi_s(\bar{\mathbf{w}}_{s-1}, \bar{\mathbf{u}}_s) \geq c \cdot (1 - p_s(\bar{\mathbf{w}}_{s-1}))$$

for all $\bar{\mathbf{w}}_{s-1}$, $\bar{\mathbf{u}}_s$, s , and i , where π_s is the propensity score

General Causal Estimand

- Average potential outcome at segment s under stochastic interventions:

$$\Psi_s(\delta_1, \dots, \delta_s) := \mathbb{E} \left[\int_{\mathcal{W}^s} Y_{is}(W_{i1} = w_1, \dots, W_{is} = w_s) \prod_{s'=1}^s dQ_{s'}(w_{s'}; \overline{\mathbf{w}}_{s'-1}, \delta_{s'}) \right],$$

where the stochastic intervention for segment s is defined as

$$dQ_s(w_s; \overline{\mathbf{w}}_{s-1}, \delta_s) = \frac{w_s \delta_s p_s(\overline{\mathbf{w}}_{s-1}) + (1 - w_s)(1 - p_s(\overline{\mathbf{w}}_{s-1}))}{\delta_s p_s(\overline{\mathbf{w}}_{s-1}) + 1 - p_s(\overline{\mathbf{w}}_{s-1})}.$$

- Identification at each segment

$$\mathbb{E}[Y_{is} \mid \overline{\mathbf{W}}_{is}, \mathbf{f}_1^{(s)}(R_{i1}), \dots, \mathbf{f}_s^{(s)}(R_{is})] = \mathbb{E}[Y_{is} \mid \overline{\mathbf{W}}_{is}, \mathbf{f}_1^{(s)}(R_{i1}), \dots, \mathbf{f}_s^{(s)}(R_{is}), \overline{\mathbf{R}}_{is}]$$

where $\mathbf{f}_{s'}^{(s)}$ is a low-dimensional **deconfounder**

- a deconfounder always exists (e.g., confounding features)
- identification works with any deconfounder
- direct adjustment of $\mathbf{R}_{is}$ leads to overlap violation

Estimation and Inference

- Apply the sequential g -formula for estimation (Robins 1986)
- Derive the influence function building on prior work (Rotnitzky et al., 2017; Kennedy, 2019)
- Additional variance due to the estimation of deconfounders
- Double-rate robustness: nuisance functions estimated at $o_p(N^{-1/4})$ rate
- Cross-fitting to estimate the asymptotic variance $\sigma_s(\delta)$
- Multiplier bootstrap for $(1 - \alpha)$ uniform confidence band for $\Psi_s(\delta)$:

$$\mathbb{P} \left(\Psi_s(\delta) \in \left[\hat{\Psi}_s(\delta) \pm \hat{c}_{1-\alpha}(\delta) \frac{\hat{\sigma}_s(\delta)}{\sqrt{N_s}} \right] \text{ for all } s = 1, \dots, s_{\max} \right) \rightarrow 1 - \alpha.$$

Super Mario Bros.TM Causal Benchmark Dataset

- Setup
 - Treatment W_{is} : appearance of Princess Peach
 - Confounder U_{is} : number of pipes (a random walk)
 - Outcome Y_{is} : number of jumps by Mario
- Confounding: Princess Peach appears more frequently when there are more pipes in each segment

$$W_{is} \sim \text{Bernoulli} \left(\epsilon + (1 - 2\epsilon) \cdot \text{logit}^{-1}(\alpha + \beta U_{is}) \right)$$

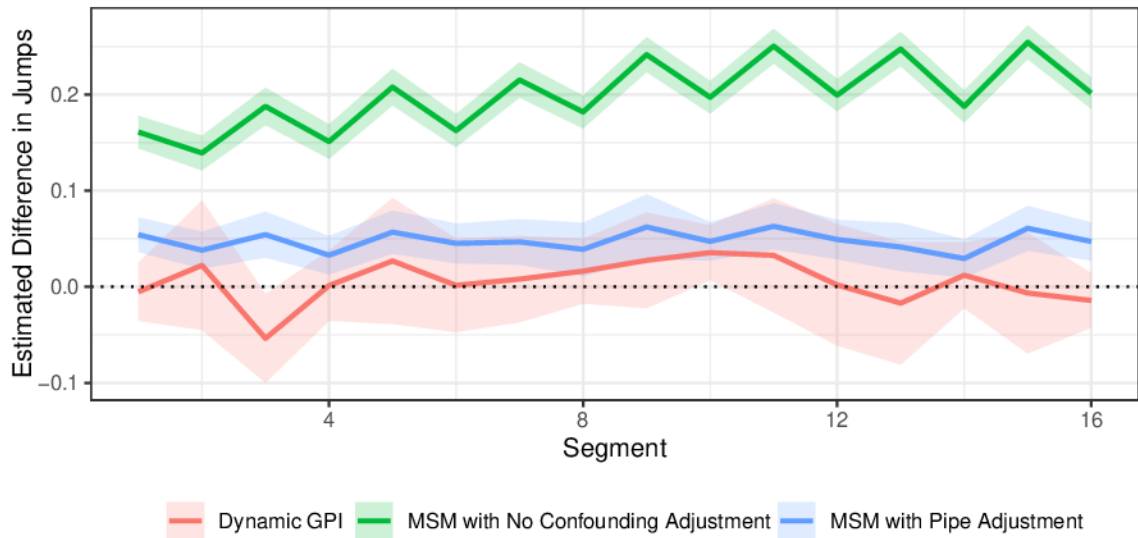
- Mario is played by a fixed reinforcement learning agent
 - realistic outcome model (complex black-box policy)
 - Princess Peach does not affect Mario's jumps
 - True causal effect is zero
- $N = 10,000$ videos with equal length and autoscroll



Analysis Setup

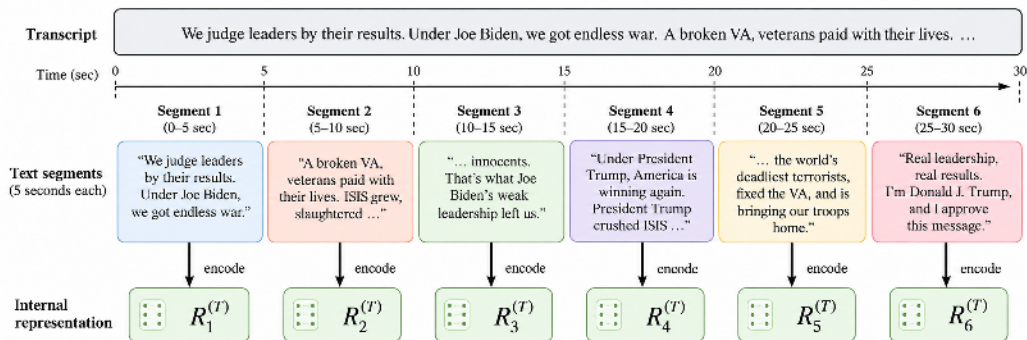
- 1-second segments
- Regenerate each video segment by **NVIDIA Cosmos Tokenizer**
 - original video dimension (per frame): $460,800 = (320 \text{ height}, 480 \text{ width}, 3 \text{ channels})$
 - dimension of internal representations: $38,400 = (40 \text{ height}, 60 \text{ width}, 16 \text{ channels})$
- Use 3-dimensional convolutional encoder in the neural network
- Estimate the trajectory of the average number of jumps
- contrast between $\delta = 5.0$ (more Peach appearance) and $\delta = 0.5$ (less Peach appearance)
- Three estimators
 - 1 **Dynamic GPI (proposed)**
 - 2 Marginal Structural Models (MSM) adjusting for the number of pipes
 - 3 Marginal Structural Models without confounding adjustment
- Same interventions and estimation strategies: only difference is confounding adjustment

Validation Results



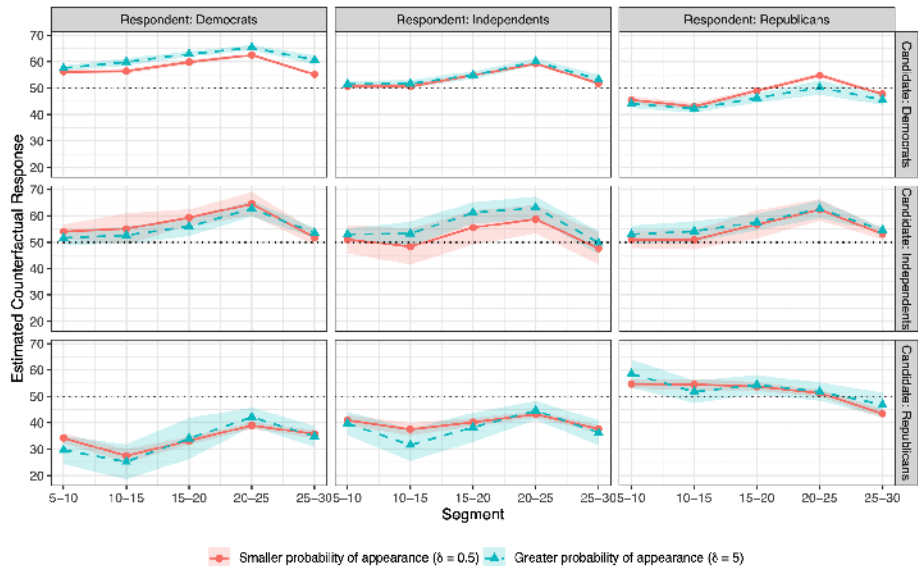
Empirical Application: Setup

- 5-second segment, regenerated by NVIDIA Cosmos tokenizer as before
- For audio data,
 - ① use Whisper for auto-transcription with time stamps
 - ② use LLaMa3.1 (8B) to regenerate transcripts and obtain internal representations
 - ③ Hidden state for the final token: dimension = 4096



- Dynamic GPI: $\delta = 0.5$ (less appearance) and $\delta = 5.0$ (more appearance)

Empirical Application: Results



Concluding Remarks

- GPI: leveraging GenAI for causal inference
 - GenAI can (re)generate high-quality videos
 - Its internal representations contain all the information in a low-dimensional format
 - We can use them to estimate the causal effects of specific video features in a dynamic setting
- Paper: <https://imai.fas.harvard.edu/research/videocausal/>
- Open-source software GPI (GenAI Powered Inference):
<https://gpi-pack.github.io/>
- Further extensions
 - better use of speech / audio information
 - discovery of treatment features
 - policy learning with unstructured treatments

