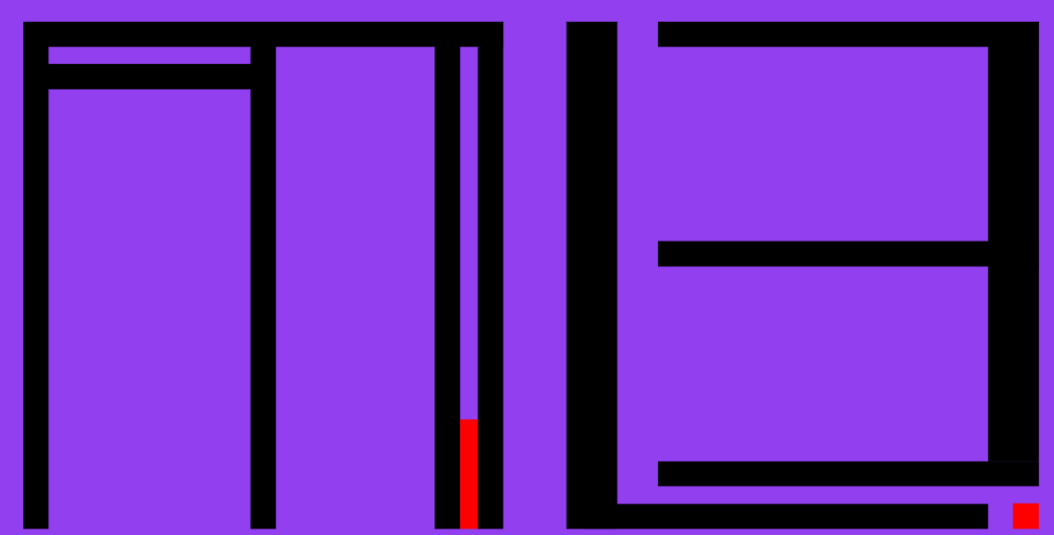


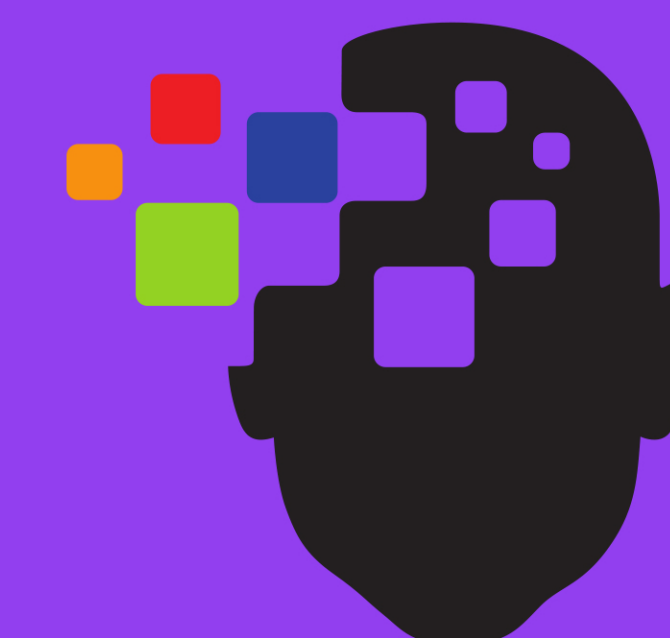


Shared Semantics, Divergent Mechanisms

Unsupervised Feature Discovery by Aligning Semantics and Mechanisms

Hyunjin Cho Youngji Roh Jaehyung Kim

Yonsei University | ICML 2026 Spotlight



ICML
International Conference
On Machine Learning

TL;DR

We jointly cluster sampled continuations by what they say and how the model computes them, recovering modes that are both semantically readable and mechanistically consistent, and causally validate each mode through targeted steering.

Motivation limits of single target analysis

One completion hides many continuation modes

Conditioning on one chosen completion can hide structure in the rest of the continuation distribution:

Semantic similarity can mislead:

similar answers, different mechanisms.

Mechanistic similarity can hide semantics:

different answers, shared mechanisms.

Single targets are brittle:

one completion may not fully represent the model.

Example: meaning and mechanism can diverge

Q: Who sings a Khmer version of "You've Got A Friend" in 1973?

A1: "...in 1973 was sung by Sok Sisowath..."

A2: "...was performed by Sok Sisowath, a notable Khmer singer..."

A3: "...was recorded in 1973 by Srey Mony..."

A1 vs. A2

Semantic Similarity: 0.95

Mechanistic Distance: 132

Similar Meaning

Different Mechanisms

A2 vs. A3

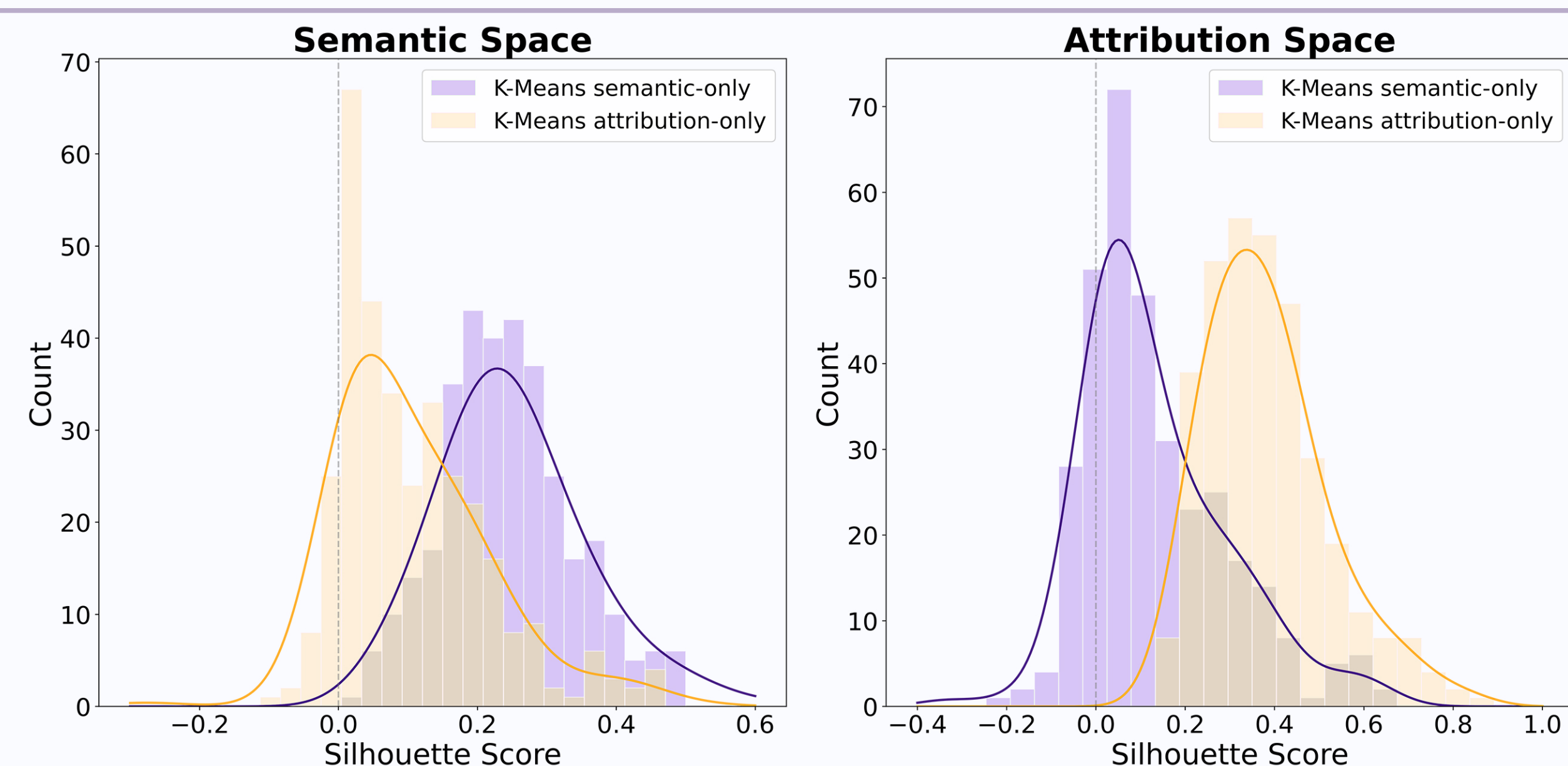
Semantic Similarity: 0.62

Mechanistic Distance: 87

Different Meaning

Similar Mechanisms

Either view alone can mislead

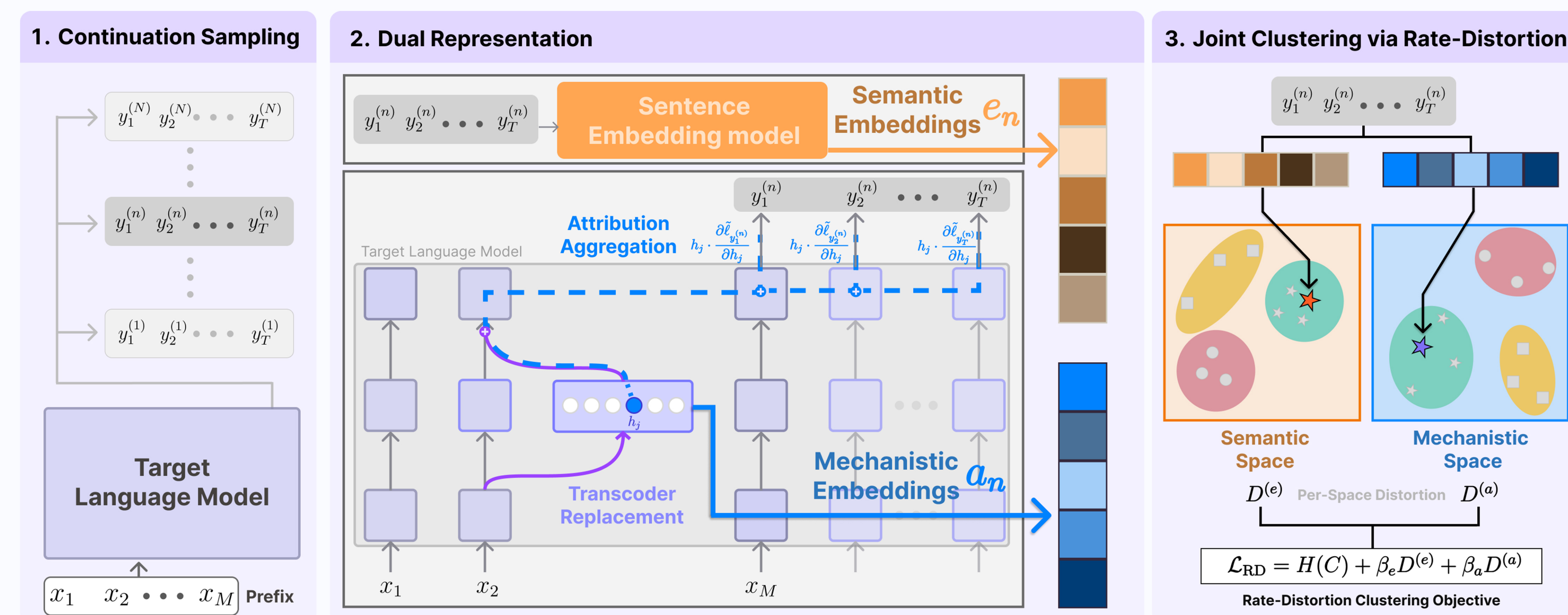


K-means learned in one view, scored in both:

clusters **clean** in semantic space,
often **poor** in attribution space, and vice versa

Method sample, attribute, cluster

Pipeline: sample, represent twice, cluster jointly



1. **Sample** many continuations per prompt, weighting by length-normalized probability.

2. **Represent** each continuation twice:

- semantic embedding (what it says)
- pooled attribution signature (which prefix features support it).

3. **Cluster** jointly via a rate-distortion objective into modes that are readable and mechanistically consistent.

Joint clustering via rate-distortion objective

Objective: $\mathcal{L}_{RD} = \underbrace{H(C)}_{\text{rate (cluster count)}} + \underbrace{\beta_e D^{(e)}}_{\text{semantic distortion}} + \underbrace{\beta_a D^{(a)}}_{\text{mechanistic distortion}}$

$$\beta_e = \gamma\beta, \quad \beta_a = (1 - \gamma)\beta$$

Assignment cost: $-\log \bar{P}_k + \beta_e \|e_n - \mu_k^{(e)}\|_2^2 + \beta_a \|a_n - \mu_k^{(a)}\|_1$

Knobs: β controls the resolution (fine vs. coarse) γ controls the balance (semantic vs. mechanistic)

Optimization: expectation-maximization with adaptive cluster count control: split / junk

Initialize $K = 1$.

E-step: assign each continuation by the combined cost above.

M-step: update semantic means, attribution weighted medians, cluster priors.

Split when distortion reduction exceeds the added rate cost.

Junk when the rate saving exceeds the distortion increase.

Repeat until convergence.

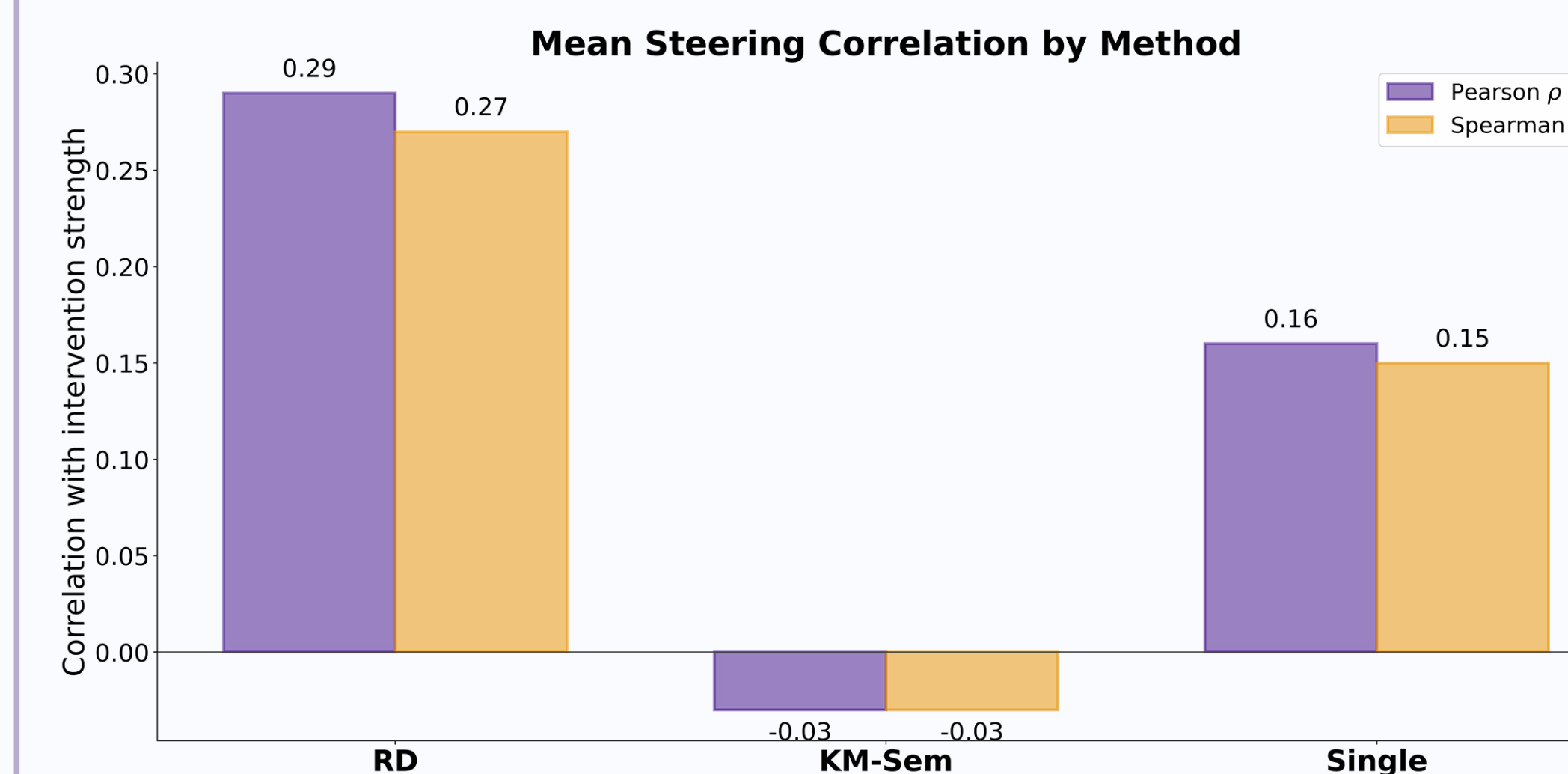
Results clusters are causally steerable

RD dominates K-means on the rate-distortion tradeoff

Method	$D^{(e)}$ semantic	$D^{(a)}$ mechanistic	D_γ mixed
KM-S	0.13	42.07	21.10
KM-A	0.18	29.48	14.83
RD (ours)	0.18	9.15	4.66

At matched rate (same number of effective clusters) RD's mixed distortion is **3–5×** lower than either K-means baseline, holding across **94%** of tested configurations. Neither view alone is competitive; only the objective consistently is.

Causal validation by steering



Q: Who pays for the renovations on Holmes Next Generation?

Source	Continuation	ΔL_{+1}	Pos.	ΔL_{-1}
C_1 :	production funding			

Held-out:

A_1 "...funded by the production company or network..."

A_2 "...HNG...funded by the production company..."

A_3 "...reality TV show...funded by the production company..."

RD-medoid "...funded through the production budget..." **+0.598 100%** **-0.189**

Single "...spinoff of the original Holmes on Homes series..." **+0.068 70%** **-0.334**

We steer along each cluster's medoid attribution direction and measure ΔL , the mean-centered shift in relative logit preference for its continuations. RD medoids give the **strongest, most consistent monotonic response**; semantic-only medoids show **near-zero correlation**. Cluster signatures are **actionable mechanistic factors, not correlational byproducts**.

Takeaway - When meaning agrees but mechanisms don't

Explaining a single chosen completion **hides heterogeneity**: the same meaning can run on different mechanisms, and vice versa. Clustering the continuation distribution jointly over both views with a **rate-distortion objective** recovers modes that are **readable, mechanistically coherent, and causally steerable**, dominating single-view baselines at matched rate.

Paper



**Project
Page**

