Attention to the Journey

Transformer Models for Customer Lifetime Value Prediction in the Presence of Marketing Interventions

Dr. Grant West · Senior Director, in4mation insights

Professor P.K. Kannan · Dean's Chair in Marketing Science, University of Maryland

Zipei Lu · Ph.D. Candidate, University of Maryland, Incoming faculty at Baruch College

February 12, 2026 · UCLA Luskin Center

The CLV Prediction Challenge

WHY CLV PREDICTION MATTERS

Customer Equity Valuation

Links customer relationships to firm valuation and investor communications

Operational Decisions

Guides prospect targeting, acquisition spending, retention investment

High-Stakes Errors

Prediction errors → wasted spend, suboptimal targeting, distorted strategy

THE FUNDAMENTAL LIMITATION

> Current methods fail to model CLV as a sequence construct or capture how a firm's own **marketing interventions** influence customer value.

The customer journey unfolds as if in a vacuum—purchases arrive according to latent parameters unaffected by:

Emails | Retargeting | Promos | Loyalty

Existing Approaches Fall Short

CUSTOMER JOURNEY AS INPUT SEQUENCE

1

Recency/Frequency Models

Fader, Hardie, Jerath (2009)

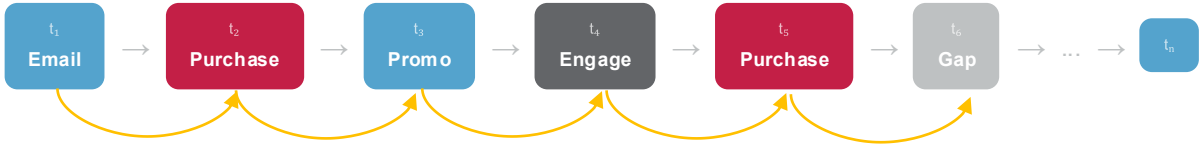
Approach: Summary statistics (recency, frequency) predict future behavior

Strength: Elegant, interpretable, minimal data

Limitation: Treats behavior as independent of firm's marketing actions

Existing Approaches Fall Short

CUSTOMER JOURNEY AS INPUT SEQUENCE



1

Recency/Frequency Models

Fader, Hardie, Jerath (2009)

Approach: Summary statistics (recency, frequency) predict future behavior

Strength: Elegant, interpretable, minimal data

Limitation: Treats behavior as independent of firm's marketing actions

2

Hidden Markov Models

Netzer, Lattin, Srinivasan (2008)

Approach: Latent states with transition dynamics

Strength: Can incorporate marketing covariates

Limitation: Markov property discards full context information, error accumulates

THE GAP

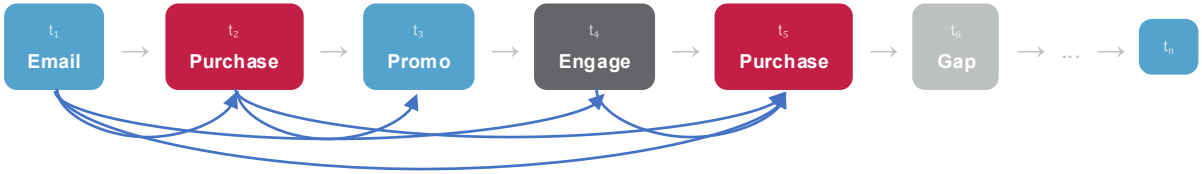
What we need:

- Learn which journey elements matter without predetermined assumptions
- Capture complex, non-linear relationships
- Handle full sequence of customer-firm interactions

Enter: Transformer Architecture

What is a Transformer?

The architecture behind GPT and modern AI breakthroughs



THE CORE INNOVATION

Current Marketing Models

MTA, MMM, RFM, Markov chains—each reduces the journey to summary statistics or assumes limited memory. Rich sequential patterns get lost.

Transformer Architecture

Sees the entire sequence at once. "Self-Attention" lets each element query all others to find relevant context—regardless of distance.

THE KEY MECHANISM: ATTENTION

Self-Attention in Plain English

For each touchpoint in a journey, the model asks: "Which other touchpoints help predict what happens next?"

How it works:

- Computes attention weights or “relevance scores" between all pairs of events, including complex interactions
- Learns these weights from data — no hard-coded rules
- Distant events and patterns stay connected if they're predictive

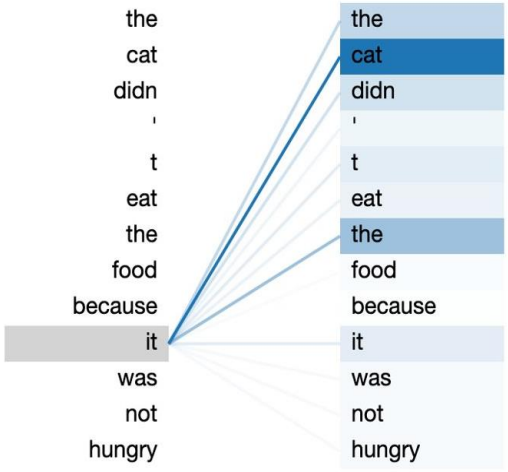
Why it matters: Transformers can discover which parts of a sequence are relevant to each other—the same capability that lets GPT understand language lets us understand customer journeys.

How the Transformer Learns from Journeys

Self-attention lets the model weigh which earlier touchpoints matter most

In Natural Language

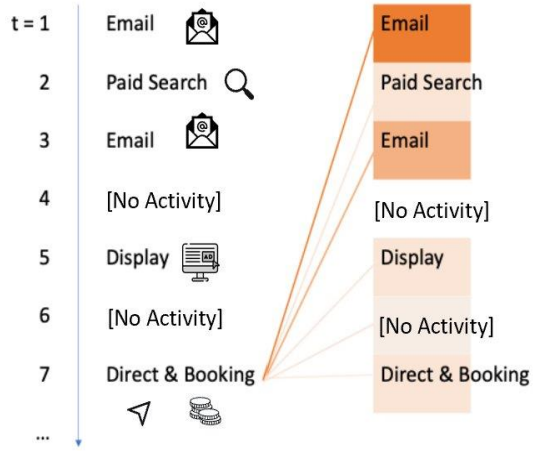
"The cat didn't eat because it was not hungry" — what does "it" refer to?



Single attention head: "it" attends most to "cat" and "the"

In Customer Journeys

Customer books directly at t=7 — which earlier touchpoints created the intent?



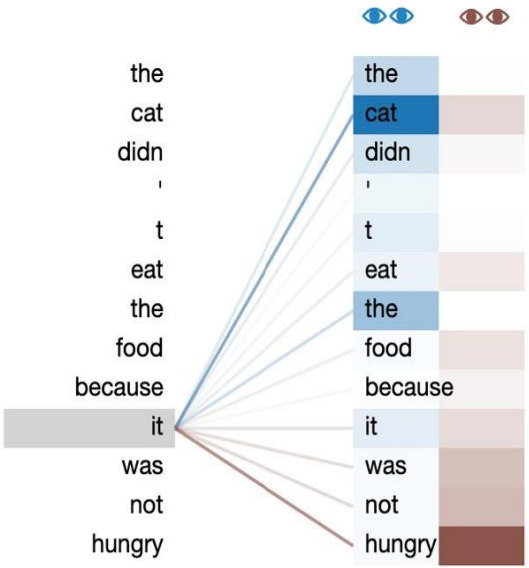
Single head: Email touchpoints get highest attention for Direct & Booking

How the Transformer Learns from Journeys

Self-attention lets the model weigh which earlier touchpoints matter most

In Natural Language

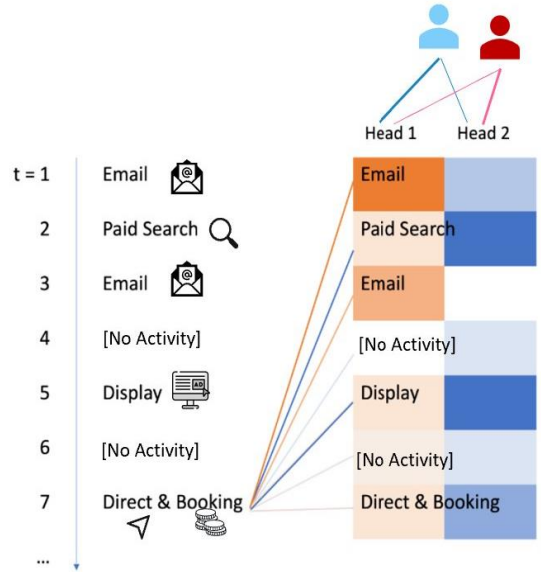
"The cat didn't eat because it was not hungry" — what does "it" refer to?



Two heads read differently: blue head → grammar, brown head → meaning

In Customer Journeys

Customer books directly at t=7 — which earlier touchpoints created the intent?



Two heads: Head 1 (orange) weighs recency, Head 2 (blue) weighs channel type

Multiple attention heads let the model read the same journey from different perspectives — one head may focus on channel type, another on timing. Mixture weights then adapt the blend to each individual customer.

Why Attention Matters for CLV

CLV is shaped by the journey itself, including marketing interventions.

A DIFFERENT WAY TO MODEL CLV

Traditional: CLV is a Latent or Stable Property

RFM assumes stable buying patterns. HMM assumes stable state transitions. Both treat CLV as something the customer "has"—not something the journey influences.

Attention: CLV is a Journey Outcome

The full sequence of customer-firm interactions—including marketing touches—shapes value.

Attention learns which patterns predict high CLV—not by rules, but by learning from thousands of actual journeys.

WHAT THIS MEANS FOR PREDICTION

Long-Range Dependencies

Attention preserves signal that other methods lose.

Complex Interactions

Attention learns conditional relationships automatically.

Different Signals, Different Weights

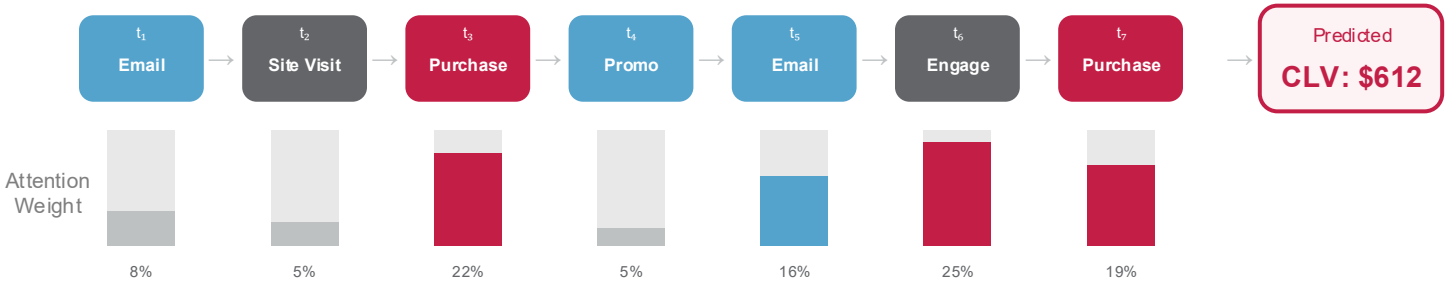
Attention weighs marketing touches, purchases, engagement, gaps by their predictive power for each customer.

← Attention weights determine which touchpoints predict value

Bottom line: Traditional CLV models quantify "how recently and frequently did they buy?" Attention quantifies "which journey patterns predict value?"—a fundamentally richer question.

What Attention Sees

Attention weights reveal which touchpoints the model relies on for CLV prediction



Key Insight

The model learns that t_6 (Engage) and t_3 (first Purchase) are strongest predictors—organic engagement after purchase signals retention. The promo at t_4 gets low attention; discount-driven behavior is less predictive of value.

Journey Heterogeneity → CLV Distribution

Different journey patterns produce a range of customer values—not just conversion vs. no conversion

C1	Email	Purchase	Engage	Email	Purchase	Referral	$1,247
C2	Ad	Site	Email	Purchase	Email	Purchase	$683
C3	Search	Ad	Site	Purchase	Gap	Purchase	$412
C4	Promo	Purchase	Promo	Purchase	Gap	Gap	$156
C5	Ad	Ad	Promo	Purchase	Gap	Churn	$47

CLV Distribution

Promo-driven, **gap**-heavy journeys **signal low CLV**

Organic engagement & **referrals signal high CLV**

Customer Lifetime Value →

Individual journeys aren't just about conversion—they predict a continuous distribution of customer value. Model-identified patterns enable better decision-making.

D2C Business Case: Customer Lifetime Value Prediction

Validation metrics for 3-month and 12-month CLV forecasting

Use Case / Dataset	Horizon	Model	MAE ($)	Pearson r	Spearman ρ	Top 10% Lift	Top 20% Lift
D2C Subscription Brand	3-mo CLV	**JWIQ**	**$43**	**0.42**	**0.59**	**3.09**	**2.50**
		RFM	$53	0.27	0.54	1.81	1.80
(same)	12-mo CLV	**JWIQ**	**$194**	**0.58**	**0.62**	**2.27**	**2.00**
		RFM	$250	0.38	0.58	2.13	1.90

−22%
MAE Reduction (12-mo)
$194 (JW) vs $250 (RFM)

+51%
Pearson r Improvement (12-mo)
0.58 (JW) vs 0.38 (RFM)

1.7×
Top 10% Lift (3-mo)
3.09 (JW) vs 1.81 (RFM)

MAE: Mean Absolute Error (lower = better) Lift: % actual CLV / % customers in decile

Client Question:

Do promotional strategies that drive short-term conversion undermine long-term value?

Conventional wisdom:

> *"Promo-acquired customers have lower lifetime value"*

The assumption: promotional incentives attract deal-seekers who churn faster and buy less.

The Naive Analysis Says Otherwise

Looking at all customers, promos appear to increase LTV

Non-Promo Acquisition

$273

12-month CLV

Promo Acquisition

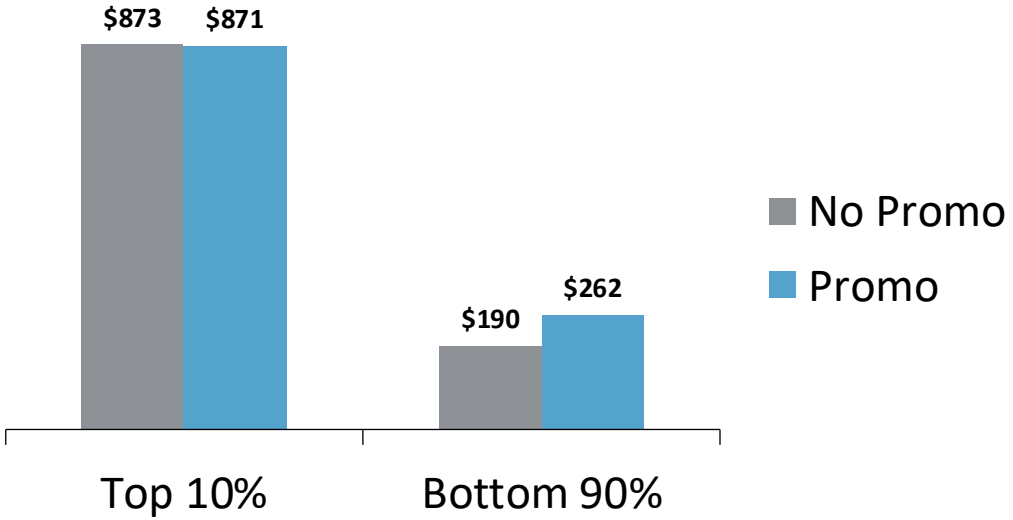
$300

12-month CLV (+10%)

At the aggregate level, promo-acquired customers are worth $27 more. *But this hides critical heterogeneity...*

But Segment-Level Tells a Different Story

The promo effect reverses for high-value customers



Top 10%: -$2.40 promo effect
Left money on the table

Bottom 90%: +$72 promo effect
Promos drove incremental customer value

KEY INSIGHT: Promos help with the broad base but leave money on the table with your best customers. *The Top 10% would have converted anyway—at full price.*

The Strategic Question

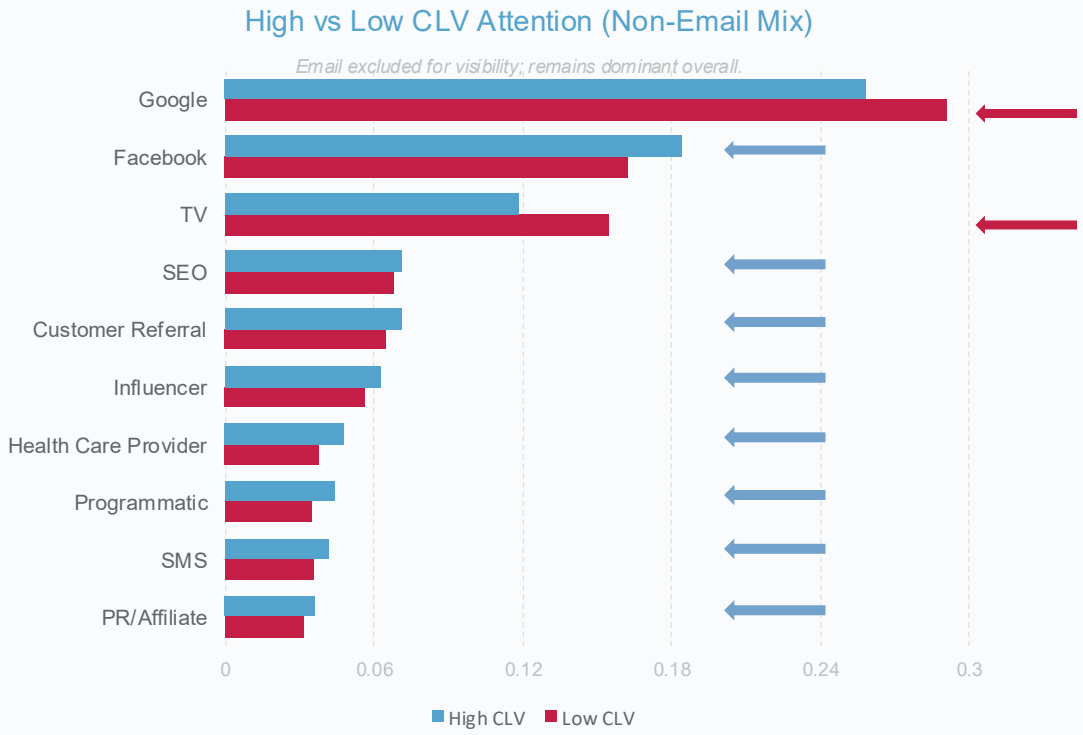
Can we identify high-value customers before the promo decision—so we know who to promote and who not to?

Withhold promos
from likely high-value converters who would purchase at full price anyway

Deploy promos
strategically to lift the Bottom 90% where promos drive incremental value

Distinctive Attention Signatures

Aggregated attention weights highlight channel and timing signals that differentiate value



KEY PATTERNS

Greater High-Value Signal

Health Care Provider, Facebook, Programmatic, Influencer, Customer Referral, SMS, PR/Affiliate

Greater Low-Value Signal

TV + Google Search lean lower-CLV; optimization opportunity

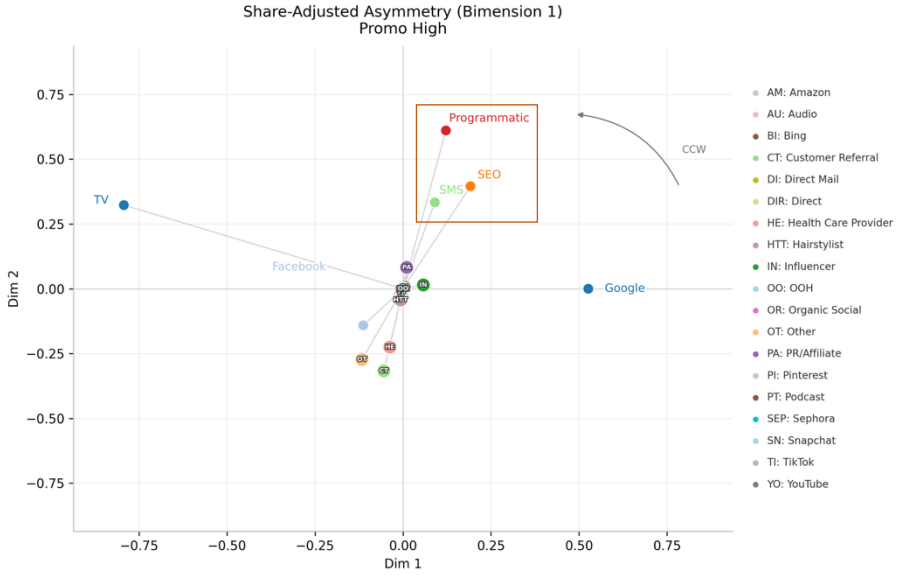
Temporal Patterns in Attention

Late-stage **~58%** > mid **~27%** > early ~14%

Potential Interpretation: The Awareness funnel (TV + Search) yields more low-value customers than channels that have an ability to target for intent.

What Does the Sequential Grammar Say?

How to interpret rotational plots depicting the results of correspondence analysis among touchpoints



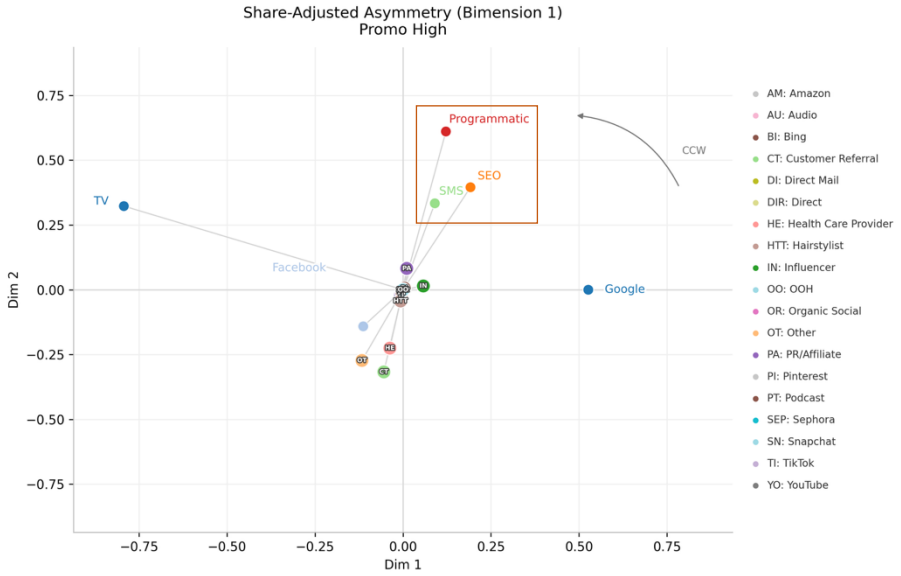
Rotational plot depicts the net directionality of transitions.

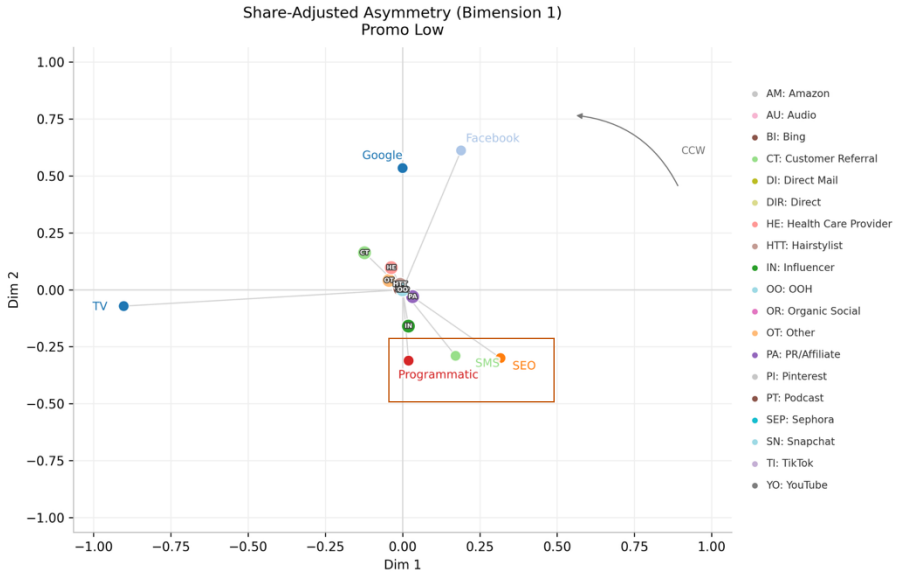
Promo High: Mid-Journey Performance Media Cluster

Performance channels (Programmatic, SEO, SMS) form a distinct mid-journey cluster above center, separated from Google. Multiple channel roles → learnable grammar.

What Does the Sequential Grammar Say?

Promo High-CLV vs. Promo Low-CLV





Promo High: Mid-Journey Performance Media Cluster

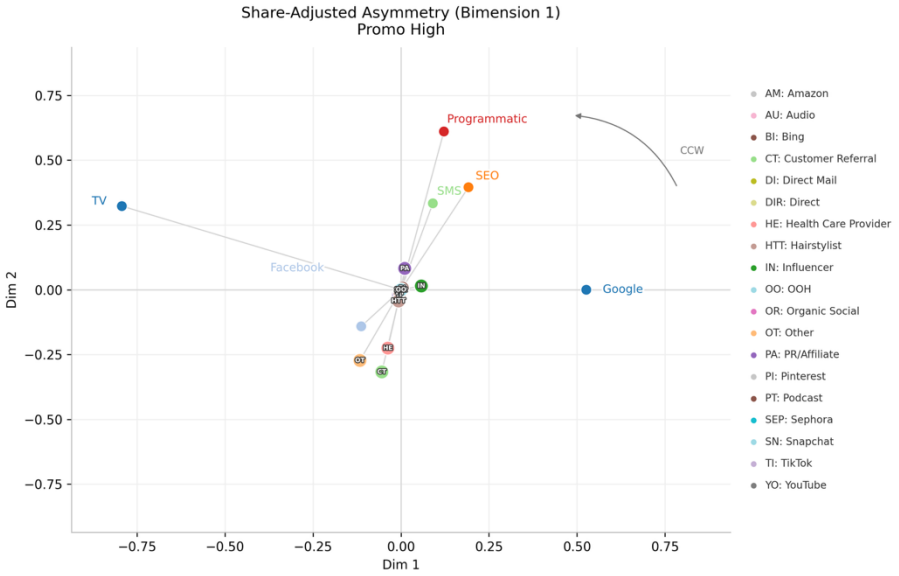
Performance channels (Programmatic, SEO, SMS) form a distinct mid-journey cluster above center, separated from Google. Multiple channel roles → learnable grammar.

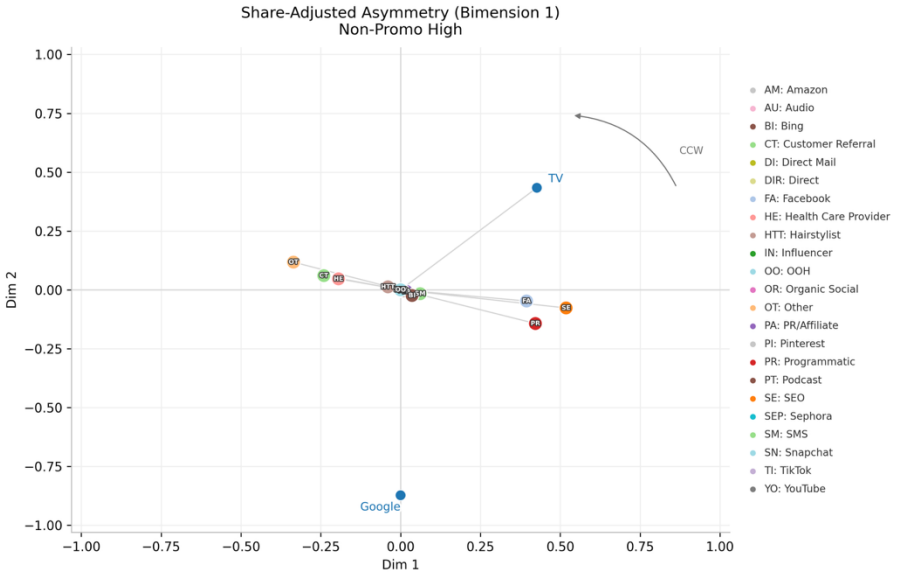
Promo Low: Displaced Performance Media Cluster

Google and Facebook rise together as single early cluster. Performance channels drop below. The sequential grammar reorganizes. Low-value journeys are more likely to begin with Facebook, while high-value journeys are more likely to begin with a Google search.

What Does the Sequential Grammar Say?

Promo High-CLV vs. Non-Promo High-CLV





Promo High: Richest sequential grammar

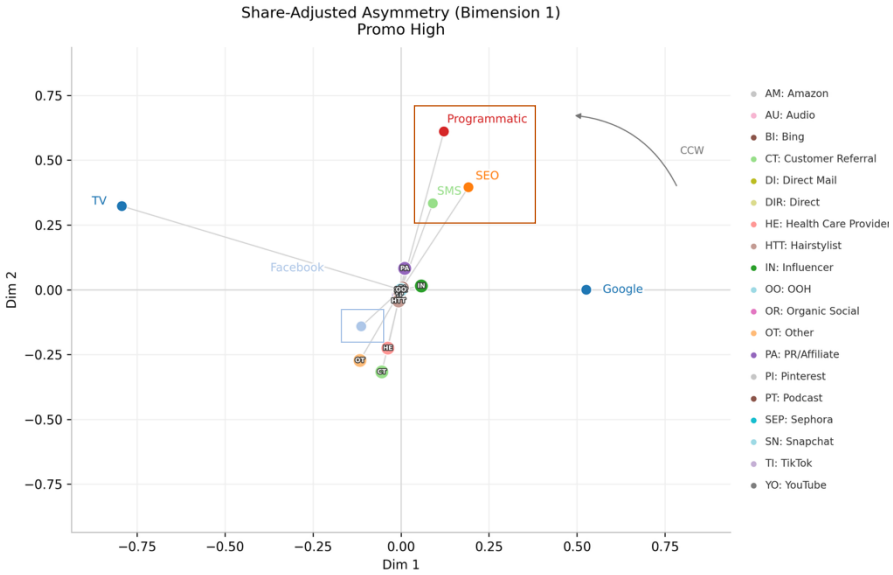
Multiple channel clusters occupy distinct positions in the rotation. Performance channels, social, and search all play differentiated sequential roles. Maximum signal for prediction.

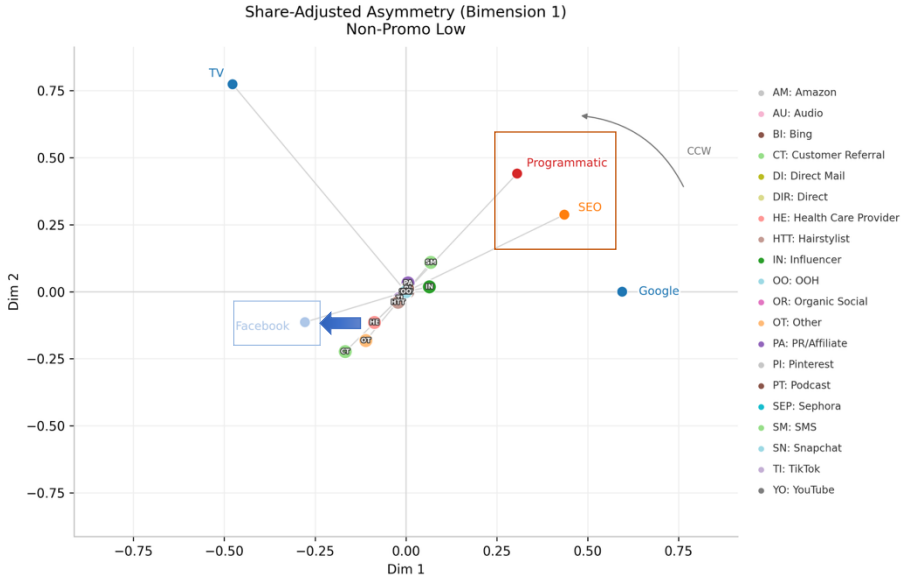
Non-Promo High: Simplest structure

Collapses to a strong polarity suggesting a classic TV-to-Google funnel movement. Everything else clusters near the origin. The journey grammar is dominated by just two channels — less nuanced structure to learn.

What Does the Sequential Grammar Say?

Promo High-CLV vs. Non-Promo Low-CLV





Promo High: Facebook = Awareness to Performance Bridge
Facebook is close to the center cluster, suggesting it plays a bridging role between the TV/awareness pole and the Google/performance pole.

Non-Promo Low: Facebook = Awareness
This is the closest to the Promo High structure. The key difference is that Facebook pulls away from the center cluster and moves further to the left, closer to TV. It behaves more like an awareness channel than a mid-funnel connector.

WHAT WE FOUND

Distinctive Attention Signatures — Promo-acquired customers show different early patterns
High-value Customers are Distinguishable — A subset shows high-value potential before promo decision
Short-term Signals Predict Long-term Value — 3-month signals visible to attention predict 12-month CLV

Insight: Promos do not destroy value, but that ***undifferentiated promo strategies*** leave money on the table. Predictive intelligence tells you ***who*** to promote—and who not to.

Implications

FOR RESEARCH

Journey Contains Predictive Signal

The customer journey itself—not merely summary statistics—contains substantial predictive information for CLV.

Complex Sequential Dependencies

Attention mechanisms reveal that customer behavior exhibits complex dependencies warranting further theoretical investigation.

Short vs Long-Term Effectiveness

Empirical evidence that promotional-CLV relationship is heterogeneous and manageable, not uniformly negative.

FOR PRACTICE

1

More Efficient Spending

Prediction improvements translate to more efficient acquisition spending and more accurate customer equity valuation.

2

Invest in Journey Data

Organizations should capture and integrate journey data - not just transactions - so they can decide who to target, when to engage, and how to align creative with channel context.

3

Actionable Insights

Attention weight interpretability offers a path to understanding which journey patterns predict high CLV.

Key Takeaways

1. **Transformer attention mechanisms substantially improve CLV prediction**
   Higher recall of high-value customers and lower prediction error are now achievable when full journey structure is modeled.
2. **Marketing touchpoints are essential to prediction accuracy**
   The customer journey - not just transaction history - contains predictive signal that can now be operationalized in CLV models.
3. **Promotional acquisition effects are heterogeneous**
   Early signals make it possible to identify high-potential customers within promotional cohorts
4. **Attention weights provide interpretable, actionable insights**
   Understanding which journey patterns predict high CLV can inform targeting strategy

Thank You

Questions and Discussion

Dr. Grant West
gwest@in4ins.com
presenter

Professor P.K. Kannan
pkannan@umd.edu

Zipei Lu
zplu@umd.edu
presenter

Lu, Z., & Kannan, P. K. (2025). AI for Customer Journeys: A Transformer Approach. *Journal of Marketing Research.*