SHOPPER: Semantic, Historical, Order-aware, and Product-conditioned Pooling for E-commerce Recommendation

AdKDD 2026 | August 10 | Jeju, Korea

Amit Jaspal
Meta Platforms, Inc.
ajaspal@meta.com

Ashkan Sadeghi
Meta Platforms, Inc.

Kevin Huang
Meta Platforms, Inc.

Nicolas Bievre
Meta Platforms, Inc.

Behnaz Poursartip
Meta Platforms, Inc.

Nikko Mizutani
Meta Platforms, Inc.

Meta

AGENDA

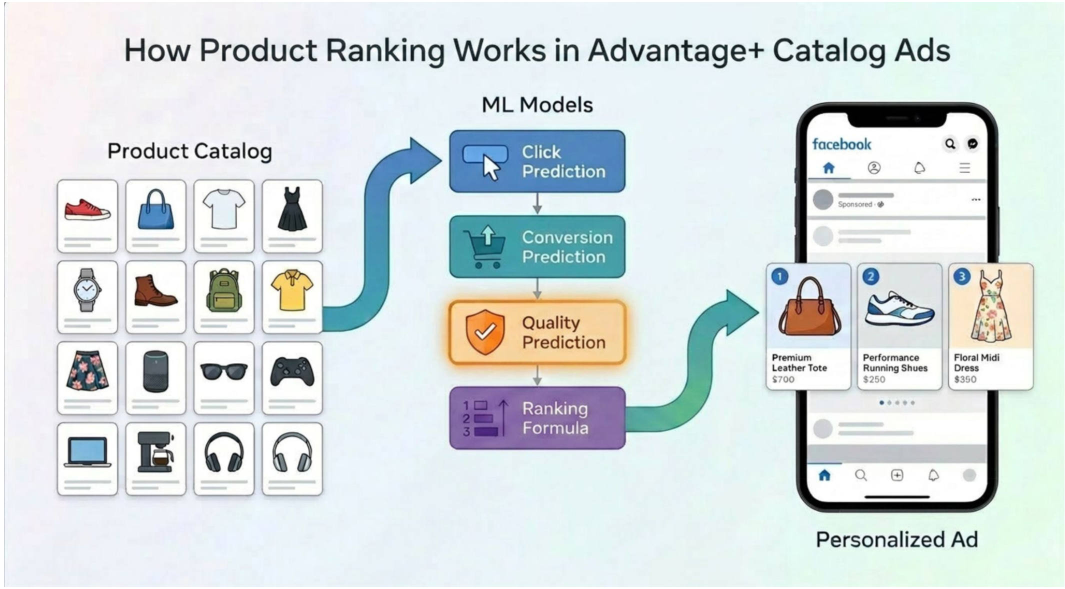
1. CONTEXT
2. PROBLEM STATEMENT
3. OUR SOLUTION: SHOPPER
4. RESULTS
5. CONCLUSION

Meta

1. CONTEXT

Product Centric Advertising (PCA) at Meta

- Dynamic ads are personalized ads where products are selected from advertiser catalogs by multi-stage and multi-objective ranking systems.
- Meta's PCA team owned these ranking models.
- Ranking must score hundreds of candidate products per impression under a strict latency budget.
- Efficient modeling of User Shopping Journey is key to unlock model improvements.

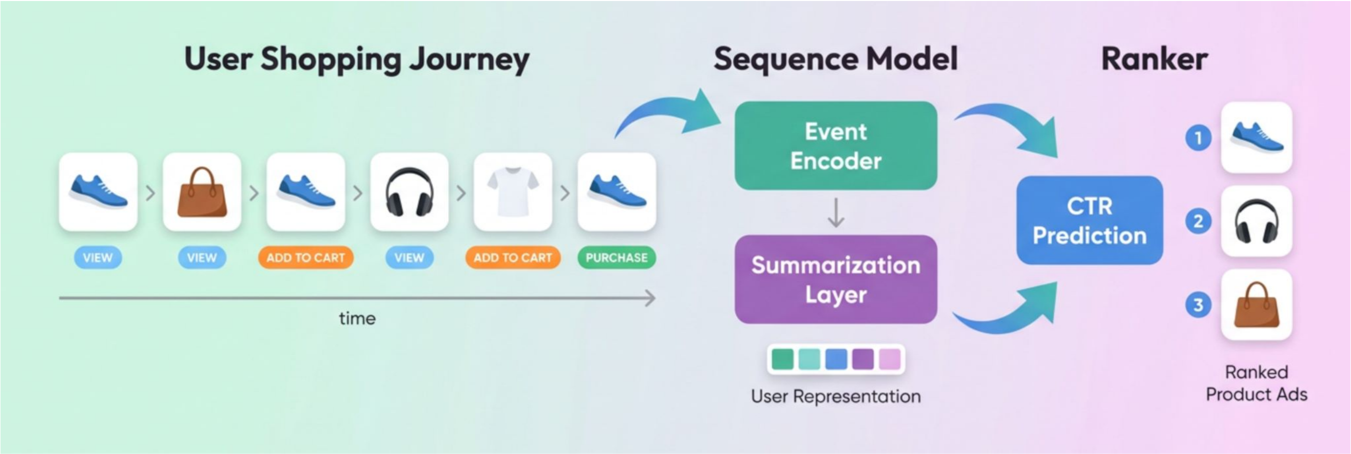


Meta

1. CONTEXT

Sequence Learning for product recommendation

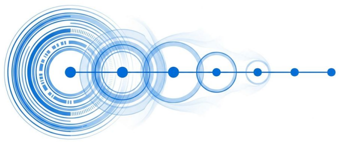
- User Shopping Journey can be naturally represented as a sequence of e-commerce events
- Most recommender systems aggregate these events into pointwise features (counts, ratios, last-N IDs) instead of leveraging their sequential nature.
- Aggregation results in the loss of event order and relationship while sequence learning keeps the journey intact.



Meta

2. PROBLEM STATEMENT

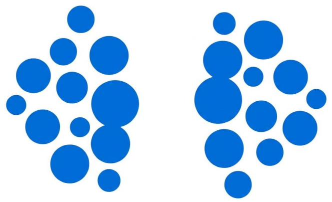
3 Key bottlenecks in Sequential Learning



Temporal Stripping

History is treated as a flat list of item IDs.

Additional context on the events are not leveraged (e.g. page type, action type, time delta between events, dwell time, position in the sequence).



Semantic Sparsity

Catalogs are enormous and fast-moving. Pure ID embeddings are excellent memorisation engines — great for retargeting inventory the user already touched. But they cannot bridge disjoint ID spaces.

Target-Agnostic

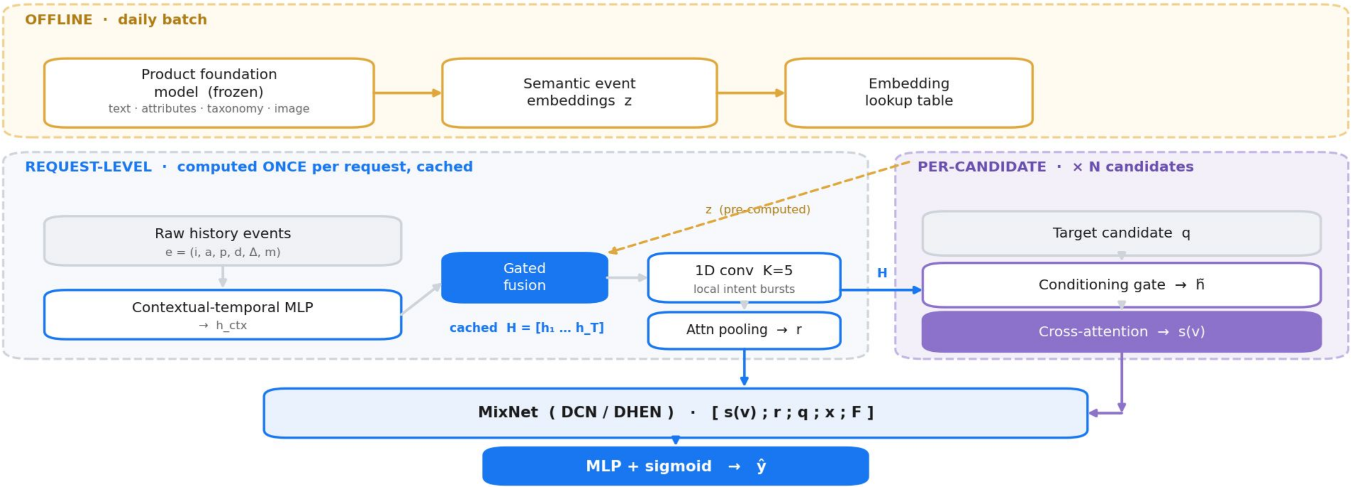
Most architectures compress the whole history into one dense vector before it ever sees the candidate. A user shopping for running shoes, a sofa and a laptop collapses into an average that describes none of them.

Meta

3. OUR SOLUTION: SHOPPER

SHOPPER Framework

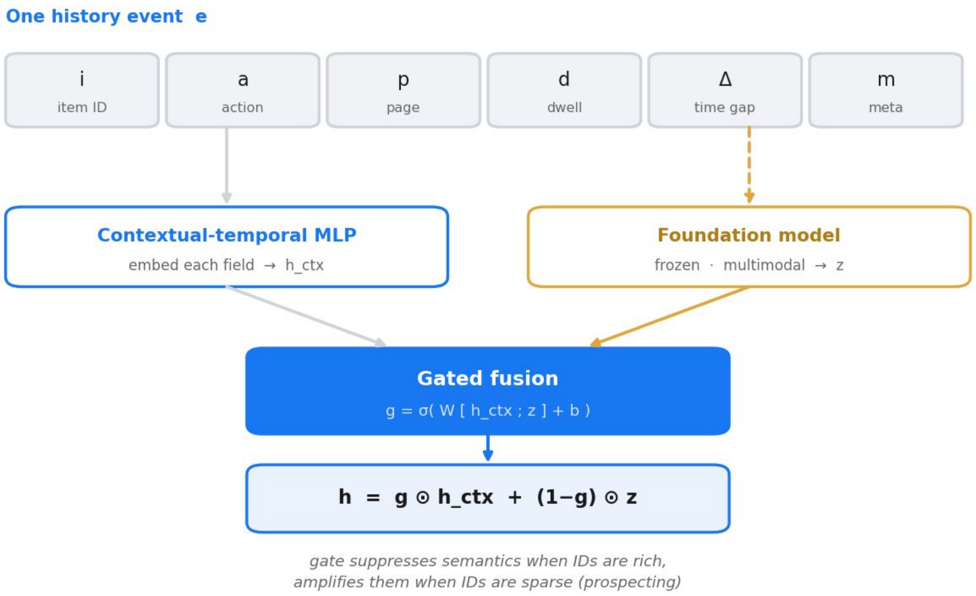
SHOPPER (Semantic, Historical, Order-aware, and Product-conditioned Pooling for E-commerce Recommendation) is a sequence modeling framework that restructures how user journeys are defined, contextualised and evaluated against a target item. Its key innovation is the factorization: expensive enrichment and local intent modeling run once per request, while target conditioning is deferred to a cheap per-candidate path.



3. OUR SOLUTION: SHOPPER

Enriched Event Representation

- Each event is a tuple e = (item ID, action type, page type, dwell time, time gap, metadata), not just an ID.
- Contextual-temporal path: every field is embedded and projected into a shared event space by an MLP.
- Semantic path: an upstream foundation model trained on product text, attributes, taxonomy and images produces z.
- A learned sigmoid gate fuses both: $h = g \odot h\_ctx + (1 - g) \odot z$.
- The gate suppresses semantics when ID signal is already rich, and amplifies it when IDs are sparse, e.g. prospecting on unseen products.

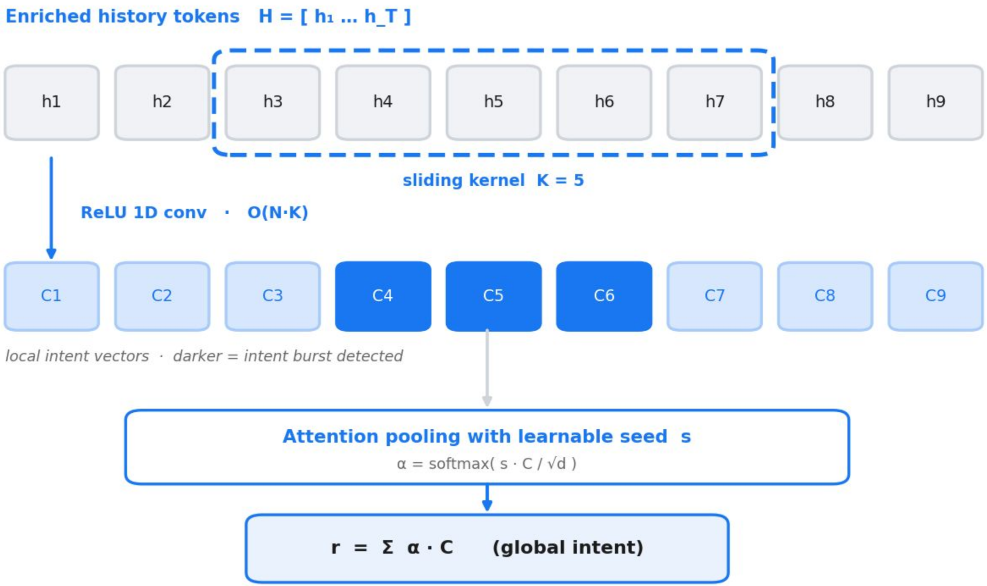


Meta

3. OUR SOLUTION: SHOPPER

Order-aware Local Intent Modeling

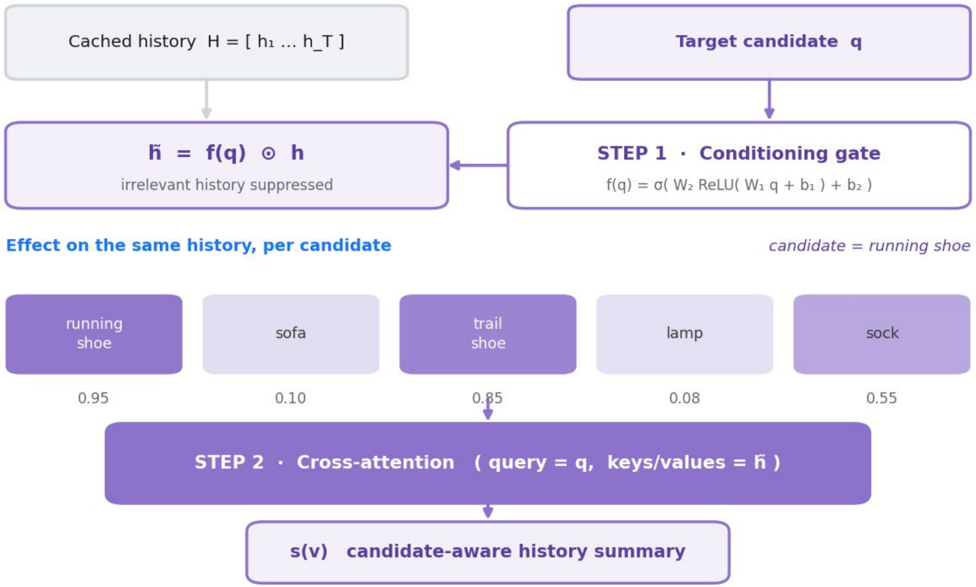
- Acute shopping intent appears as short bursts of related actions, not isolated clicks.
- A 1D convolution over the enriched tokens captures those local patterns at O(N·K) cost, avoiding quadratic self-attention.
- Kernel size K = 5, selected by grid search over {3, 5, 7, 9}.
- A learnable seed vector attends over the convolved sequence to produce the global intent vector r.
- All of this is request-level: computed once and reused across every candidate.



3. OUR SOLUTION: SHOPPER

Product-conditioned Pooling

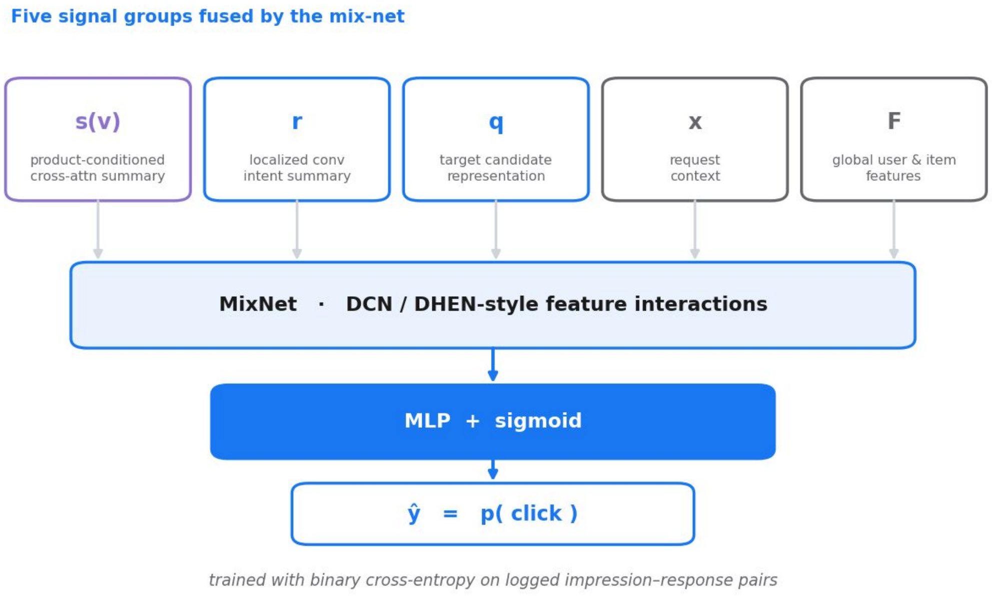
- Summarizing history before seeing the candidate averages away the evidence that matters.
- Step 1 (conditioning gate): $f(q) = \sigma(W_2 \mathrm{ReLU}(W_1 q + b_1) + b_2)$, applied element wise as $\tilde{h} = f(q) \odot h$.
- Step 2 (cross-attention): the target acts as the query and summarizes the conditioned sequence into s(v).
- The target does not merely reweight a precomputed vector; it shapes which parts of the history are extracted.
- Only these two lightweight operations run per candidate.



3. OUR SOLUTION: SHOPPER

Final Scoring and Optimization

- Five signal groups are fused by the mix-net: s(v), r, q, request context x, and global user / item features F.
- MixNet applies DCN / DHEN-style feature interactions; a final MLP with sigmoid produces ŷ.
- Trained as a ranking task on logged impression–response pairs with binary cross-entropy loss.
- Normalized Entropy (NE) is the primary metric. On internal datasets a 0.02% NE gain is already significant.



Meta

3. OUR SOLUTION: SHOPPER

Serving Optimization

Three tiers of computation keep deep target-aware sequence modeling inside the production latency budget.

OFFLINE

1× per day

Foundation-model embeddings pre-computed by a daily batch pipeline, served through lookup

REQUEST-LEVEL

1× per request

Event enrichment + 1D conv intent modeling. H and r are cached and reused by all candidates

PER-CANDIDATE

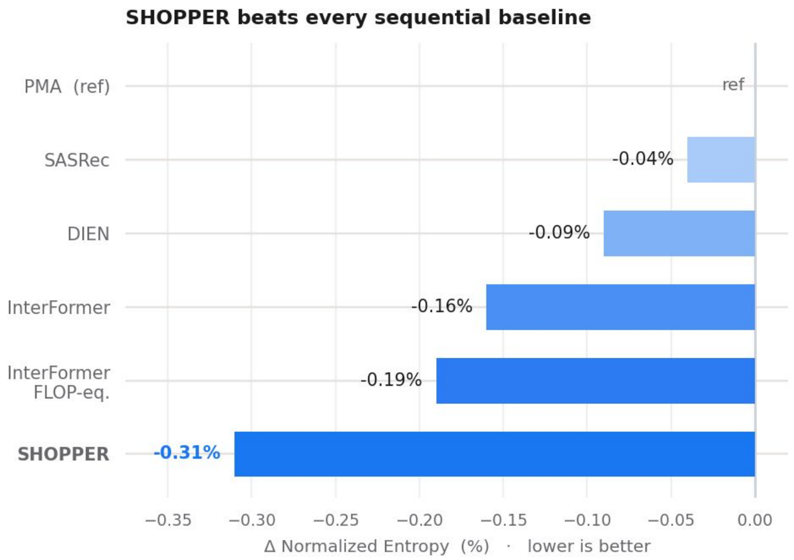
N× per request

Only the conditioning gate and cross-attention run per item — two lightweight cached-tensor ops

Deep target-aware sequence modeling in production — no p99 latency regression vs baseline

Offline Results: SHOPPER vs Baselines

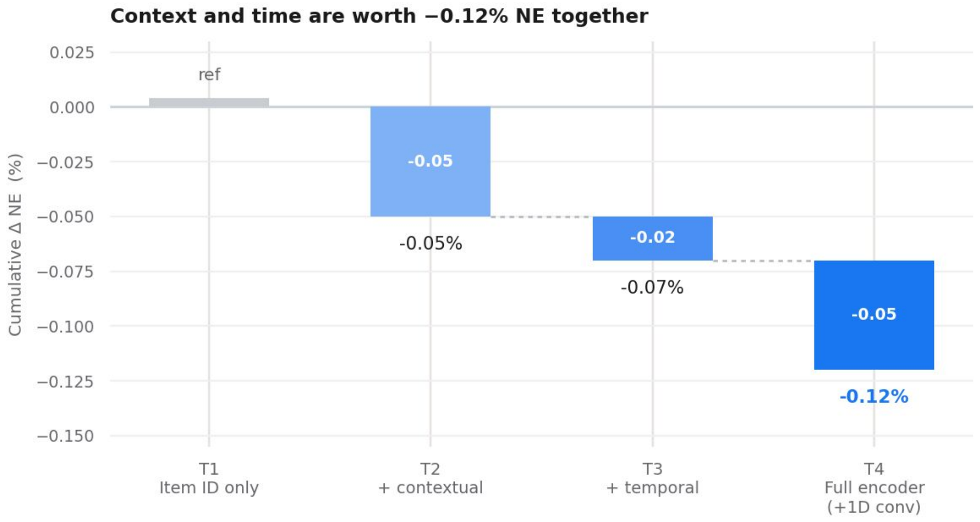
- Billions of logged impression–click pairs from a large-scale proprietary e-commerce dataset.
- All baselines use comparable DCN / DHEN mix-nets; a FLOP-equivalent InterFormer controls for compute budget.
- SHOPPER reaches −0.31% NE against the PMA reference, a wide margin where 0.02% is significant.
- It beats both target-agnostic (PMA, SASRec) and target-aware (DIEN, InterFormer) families, so the gain comes from the combination rather than any single part.



4. RESULTS

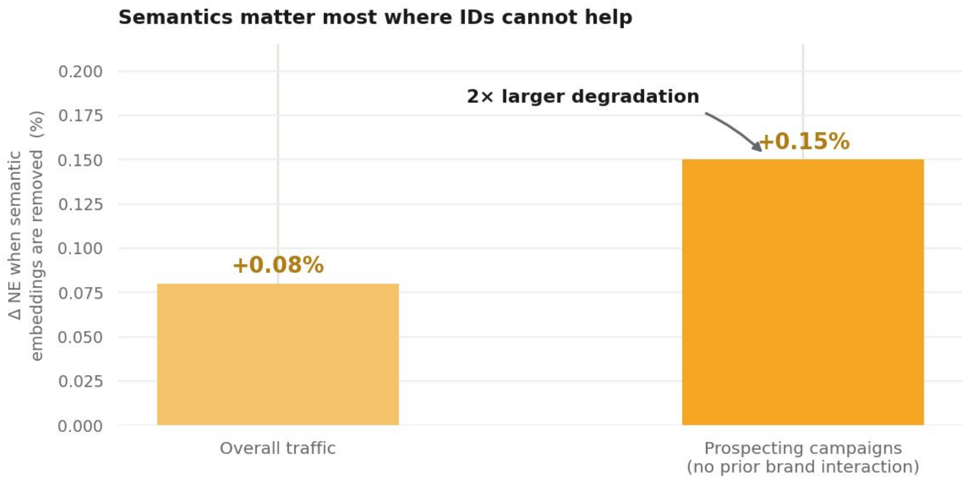
Ablation: Contextual and Temporal Encoding (RQ-2)

- Contextual features (action type, page type) alone: −0.05% NE.
- Temporal features (timestamp deltas, dwell time): a further −0.02% NE.
- The 1D convolutional intent pooling adds another −0.05% NE on top.
- Full encoder: −0.12% NE versus an item-ID-only baseline.
- The convolution gain proves intent lives in multi-event temporal patterns that per-event features cannot represent.



Ablation: Semantic Enrichment (RQ-3)

- Removing foundation-model embeddings degrades overall NE by +0.08%.
- Stratified by campaign objective, prospecting degrades by +0.15%, roughly 2× worse.
- Prospecting is exactly where users have no prior interaction with the target brand.
- This confirms the semantic-sparsity bottleneck: ID-only models fail precisely where cross-catalog generalization is needed most.



Online Results

A two-week A/B test on live e-commerce advertising traffic confirmed the offline findings, and stayed inside the latency budget.

+0.4%

top-line business metric

statistically significant

2-week live A/B

on production ranking traffic

No p99 latency regression

vs the production baseline

Three system optimizations that made deployment possible

1

Foundation-model embeddings pre-computed offline daily and served via lookup

2

Enriched events + 1D conv run once per request and cached across all candidates

3

Product-conditioned pooling and cross-attention optimized for GPU-accelerated inference

5. CONCLUSION & FUTURE WORK

Limitations and Future Work

Known operational limits of the current system, and where SHOPPER goes next.

OPERATIONAL LIMITATIONS

Embedding staleness

Semantic quality depends on how often the upstream foundation model is refreshed — fast-moving catalogs can drift.

Fixed sequence length

A hard maximum length discards older but still-informative history signals.

Cold start for new items

Offline pre-computation leaves newly created products without vectors at serving time.

FUTURE WORK

Incremental embedding updates

Refresh semantic representations continuously instead of in a daily batch, closing both the staleness and cold-start gaps.

Adaptive-length sequence modeling

Let the model decide how much history to keep per user rather than truncating at a fixed T.

Broader multimodal signals

Extend the enrichment path with richer image and cross-surface engagement cues.

SHOPPER resolves the quality-vs-latency trade-off — the remaining work is keeping semantics fresh

Q & A

Corresponding Author

Amit Jaspal
Meta Platforms, Inc.
ajaspal@meta.com

Meta