Two-Tier Visual Triggering for Adaptive Shopping Recommendations at Pinterest

Kshiteesh Hegde✉
Pinterest, Inc.
San Francisco, CA, USA
khegde@pinterest.com

Ai Zhang
Pinterest, Inc.
San Francisco, CA, USA
aizhang@pinterest.com

Kayoung Lee
Pinterest, Inc.
San Francisco, CA, USA
kayounglee@pinterest.com

Weiran Li
Pinterest, Inc.
San Francisco, CA, USA
wli@pinterest.com

Sai Xiao
Pinterest, Inc.
San Francisco, CA, USA
sxiao@pinterest.com

Janet Ryu
Pinterest, Inc.
San Francisco, CA, USA
jryu@pinterest.com

Abstract

Adaptive recommender systems often rely on trigger logic that decides whether a contextual module should surface for a given item. In visual discovery platforms where users move from inspiration to action, this trigger decision is especially challenging: a shopping module should appear when a user is close to realizing an aspirational journey, yet there are opportunities to improve both coverage: surfacing the module on eligible content that was previously unreached, and precision: ensuring the module appears only where shopping is the natural next step. We present a two-tier visual triggering system deployed at Pinterest for an adaptive shopping module. The first tier expands recall through an upgraded content interest classifier and broader category coverage. The second tier improves precision by clustering candidate images using visual embeddings, labeling representative cluster centroids with a large language model to assess visual shopping-worthiness, and propagating eligibility decisions to cluster members via low-latency key-value lookup. This design enables rich visual understanding at labeling time while keeping the serving path free of online model inference. A human evaluation of 1,000 random samples showed 91% agreement between model labels and human judgment. An online A/B experiment demonstrated a +11.2% lift in module engagement (saves) and a +1.0% lift in site-wide saves, confirming that better trigger quality improves broad user engagement, along with module-local metrics. We describe the architecture, the iterative experiment design across multiple versions, trade-offs in cluster granularity, and practical lessons for building adaptive triggers in production visual recommender systems.

CCS Concepts

• **Information systems** → **Recommender systems**; *Retrieval models and ranking*; • **Applied computing** → **Online shopping**; • **Computing methodologies** → Computer vision.

OARS @ RecSys '26, Minneapolis, MN

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Keywords

adaptive recommender systems, trigger eligibility, multimodal recommendation, cluster labeling, visual language models, online experimentation, visual discovery, shopping recommendation

ACM Reference Format:
Kshiteesh Hegde, Ai Zhang, Kayoung Lee, Weiran Li, Sai Xiao, and Janet Ryu. 2026. Two-Tier Visual Triggering for Adaptive Shopping Recommendations at Pinterest. In *Proceedings of Online and Adaptive Recommender System at RecSys (OARS @ RecSys '26).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Visual discovery platforms enable users to move from inspiration to action: collecting ideas, comparing alternatives, and eventually finding concrete products that bring an idea to life. Pinterest, with over 600 million monthly active users, exemplifies this journey. Users save and organize visual content (Pins) into collections (boards), and the platform provides increasingly sophisticated tools to bridge the gap between aspiration and purchase. One such tool is the Shop the look (STL) module (see Figure 1), an adaptive shopping surface that appears when users open Pins (Closeup) and the system determines that the content is shopping-oriented. STL uses visual object detection to identify individual items within an image (e.g., a jacket, a pair of shoes, a table lamp) and presents visually similar shoppable product recommendations for each detected object [8].

The effectiveness of such a module depends critically on a trigger-eligibility decision: for a given image, should the shopping module be surfaced at all? A module that appears on the wrong content damages user trust and wastes valuable screen real estate that could be allocated to other recommendation surfaces. Conversely, a module that fails to appear on highly shoppable content represents a missed opportunity to connect users with products they are ready to engage with. This binary decision is simple in formulation yet complex in practice because visual content spans an enormous range of domains, styles, and user intents.

The prior trigger logic for STL relied on two checks: (1) whether the image belonged to one of two eligible content categories (fashion and home decor) as determined by a content interest classifier, and (2) whether the image contained detectable shoppable objects. This approach served as a strong baseline and enabled the module to reach millions of users. However, analysis revealed an opportunity to substantially improve trigger quality in two directions.

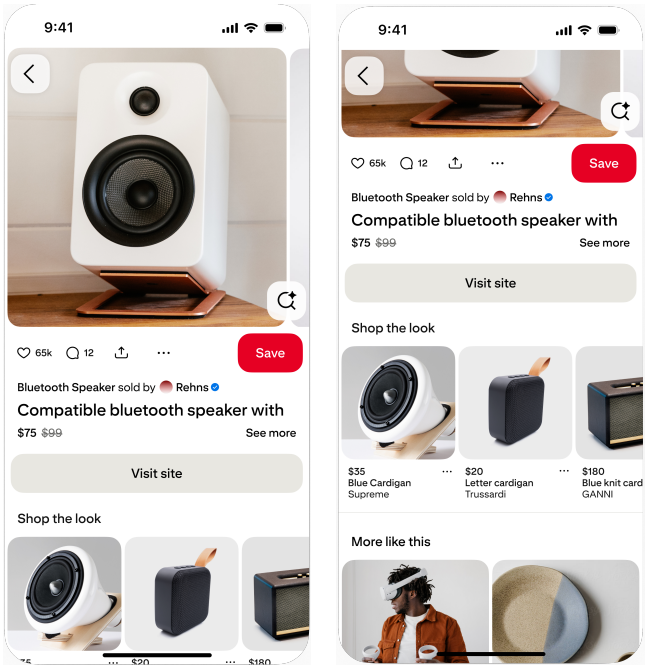

Figure 1: Example of Pinterest's Shop the look (STL) module. The module presents a horizontally scrollable set of visually related, shoppable product recommendations beneath the primary Pin, enabling users to discover and navigate to similar or complementary items.

First, the module could be extended to additional high-intent shopping verticals and other visually rich domains that were not yet covered by the four original categories. Second, precision could be improved by better distinguishing images where shopping is the natural next step from those where the visual content, while categorically eligible, does not lend itself to a product-exploration experience.

In this paper, we present a two-tier visual triggering system that decomposes the eligibility decision into complementary layers. The first tier (T0) addresses recall by expanding category coverage and upgrading the underlying interest classifier. The second tier (T1) addresses precision by clustering candidate images in a visual embedding space, labeling representative cluster centroids with a large language model (LLM) to assess visual shopping-worthiness, and propagating eligibility labels to all cluster members through a lightweight online lookup. The design was inspired by recent work on cluster-level generative labeling for multimodal search [9], which demonstrated that labeling representative images at the cluster level can generalize effectively to cluster members. We adapt and extend this principle to the distinct problem of binary trigger-eligibility classification, building a purpose-built architecture that addresses the unique challenges of module gating: stricter accuracy requirements (a wrong trigger inserts or removes an entire experience), interaction with other recommendation surfaces, and the need for iterative tuning through sequential online experimentation. A pictorial illustration of the architecture is shown in Figure 2.

Our contributions are: (1) We formulate module triggering as a two-tier problem that separates recall (rule-based category expansion) from precision (LLM-assisted cluster-level visual eligibility), enabling independent tuning of each layer. (2) We design and deploy a production system that leverages offline LLM-based visual assessment at the cluster level while keeping online serving free of model inference. (3) We report results from a production A/B experiment across multiple iterative versions, culminating in +11.2% module saves, +5.% long clicks, and +1.0% site-wide saves. (4) We discuss practical lessons around iterative experiment design, cluster granularity, embedding selection, and when cluster-level eligibility is an appropriate design pattern for adaptive modules.

2 Related Work

2.1 Triggering and Gating in Recommender Systems

Adaptive recommender systems must decide not only what to recommend but also when and where to present a recommendation surface. This gating decision determines whether a recommender has any opportunity to serve the user. In e-commerce, trigger logic governs when product carousels, comparison tools, or visual try-on features appear. In content platforms, it governs when contextual modules like "related products" or "shop this look" are activated. Despite its importance, trigger logic is often implemented as a collection of hand-crafted rules that accumulate technical debt over time as content distributions shift and new verticals emerge [7].

Recent work has explored learned triggers using engagement prediction [2], contextual bandits [4], and reinforcement learning [10]. However, these approaches often require large amounts of interaction data for cold-start surfaces, add latency through online model inference, and can be difficult to inspect and debug. Our system offers an intermediate approach: the triggering logic incorporates LLM-based visual judgments at scale but avoids online inference by precomputing decisions offline and serving them through key-value lookup.

2.2 Cluster-Level Labeling for Visual Systems

The idea of labeling representative cluster centroids and propagating labels to members has been explored in several contexts. Zhang et al. [9] introduced Generative Labeling On Clusters (GLOC) for scalable multimodal image insight generation, showing that cluster-level labeling with a vision-language model can produce high-quality image descriptors and query suggestions at a fraction of per-image labeling cost. Their work achieved a 2000x reduction in labeling volume and significant online engagement gains for a multimodal search application at Pinterest. Li et al. [5] described the broader multi-modal search system that consumes these cluster-level labels as pivot suggestions and descriptors.

Our work was inspired by the cluster-level labeling principle demonstrated in GLOC, but addresses a fundamentally different problem. Where GLOC generates open-ended textual insights (descriptors, query pivots), our system produces binary eligibility decisions for module gating. This difference has several implications:

(a) the accuracy bar is higher because a wrong trigger inserts or removes an entire user experience rather than showing a suboptimal text label; (b) the system must interact with a rule-based recall tier that controls the candidate pool; (c) the evaluation requires iterative online experimentation to understand interactions with other recommendation surfaces; and (d) the prompt design, embedding selection, and cluster granularity must all be tuned for a binary classification objective rather than open-ended generation. We built a dedicated pipeline addressing these requirements rather than extending the existing infrastructure.

2.3 Visual Embeddings for Content Understanding

Large-scale visual similarity clustering has been widely used in recommendation and search systems for tasks such as large-scale image retrieval, visual similarity search, and product discovery [1, 6]. The key insight that makes cluster-level labeling viable is that high-quality visual embeddings produce clusters where members genuinely share semantic attributes relevant to downstream tasks [7]. Visual embedding models trained on billions of user interactions can provide strong alignment between visual similarity and user-perceived relevance, making them well-suited for propagating eligibility judgments across visually similar images.

3 Problem Statement

The STL module surfaces shoppable product recommendations on Pin Closeup. When a user taps into a Pin, the module identifies shoppable objects using visual object detection and retrieves visually similar products from a merchant catalog. The module is valuable when the image depicts product-oriented content in a contextually appropriate setting—for example, an outfit photo where the user might want to find similar clothing items, or a living room scene where the user might want to purchase similar furniture.

The prior trigger logic operated as follows: (1) compute a content interest score across fashion and home decor; (2) verify that the image contains at least one detectable shoppable object. If both conditions were met, the module was shown. This logic provided a solid foundation, but our analysis identified two opportunities for improvement:

3.1 Precision opportunity

Some images that passed category checks represented contexts where shopping was not the natural next action for users. Improving precision for these cases could increase per-impression engagement quality.

3.2 Coverage opportunity

Content in categories outside fashion and home decor was not yet eligible, regardless of shopping intent. Wedding Pins (table settings, floral arrangements, venue styling) represent highly shoppable content with strong user interest. Similarly, DIY & Craft Pins often contain purchasable supplies and tools. Extending coverage to these verticals could unlock significant new engagement.

The redesign goals were: (1) improve recall by capturing shopping-eligible content in underserved categories, (2) improve precision by filtering out content that passes category checks but is not visually suitable for shopping, (3) avoid recurring costs of per-image online model inference, and (4) deliver measurable improvement in both module engagement and sitewide user satisfaction metrics.

4 System Design

4.1 Architecture Overview

Our system decomposes the trigger-eligibility decision into two complementary tiers, each targeting a distinct improvement opportunity. Tier 0 (T0) focuses on recall: it broadens the candidate pool by expanding category coverage and upgrading the content interest signal. Tier 1 (T1) focuses on precision: it applies cluster-level visual eligibility labels to filter the expanded pool down to content where the shopping module is genuinely appropriate. The two tiers are designed to be independently tunable—T0 parameters (category set, score thresholds, exclusion rules) can be adjusted without rebuilding T1, and vice versa.

At serving time, the decision path for a given image is: (1) retrieve the content interest scores across eligible categories; (2) apply T0 filtering rules (category eligibility, score threshold, exclusion list); (3) if T0 passes, look up the image's cluster membership and retrieve the precomputed T1 eligibility label; (4) if the T1 label is "eligible" and the image contains detectable shoppable objects, surface the module. Steps 1–4 involve only cached scores and key-value lookups—no online model inference is required.

4.2 Tier 0: Rule-Based Recall Expansion

The first tier expands coverage by broadening the set of content considered for the module. Three changes were made to the rule-based layer:

4.2.1 Classifier upgrade. The content interest classifier was upgraded to a newer version, which provides a stronger and better-calibrated signal across content categories. The improved classifier produces fewer boundary misclassifications in ambiguous domains.

4.2.2 Category expansion. Two additional categories—weddings and DIY/craft—were added to the eligible set. These categories were selected based on offline analysis showing high shopping intent but zero module coverage under the prior logic. Weddings, in particular, represent a major shopping vertical where users actively seek purchasable items (decor, attire, accessories) but were previously invisible to the trigger.

4.2.3 Negative category filtering. Images whose dominant content category is clearly non-shoppable are explicitly excluded. These categories frequently contain visually rich images that pass interest thresholds but represent contexts where shopping recommendations are inappropriate.

These T0 changes are intentionally conservative: they expand recall by adding known-good categories and filtering known-bad ones, without requiring expensive labeling or model training. The expanded candidate pool is then passed to T1 for precision refinement.

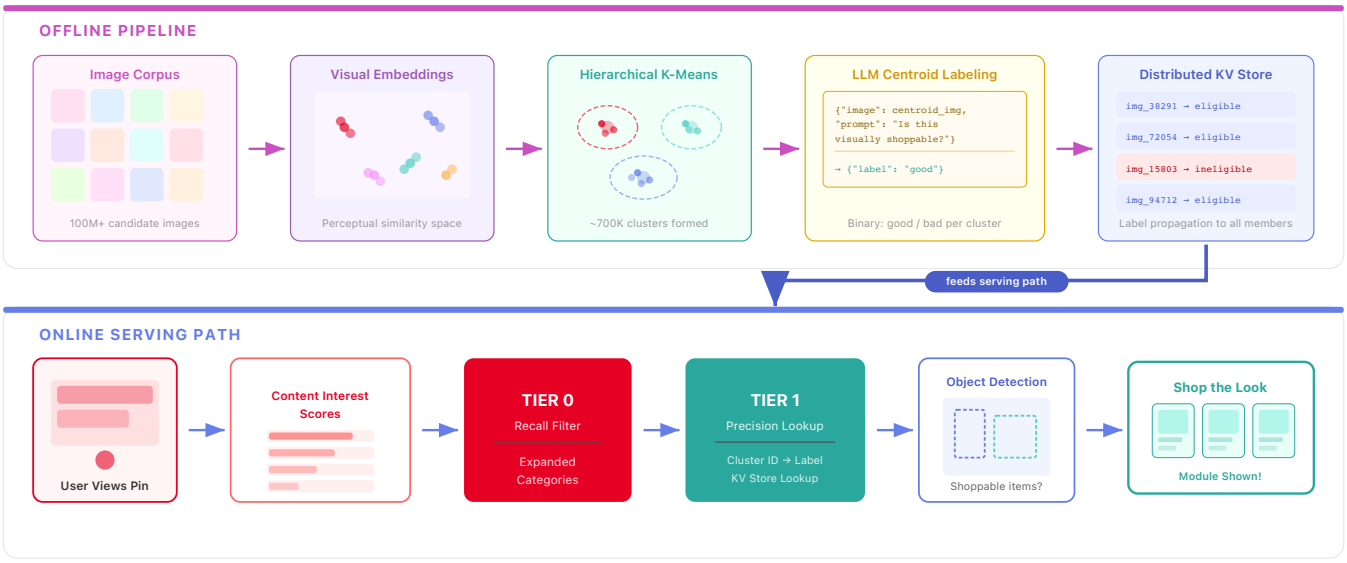

Figure 2: Overview of the two-tier visual triggering system. The offline pipeline (top) computes visual embeddings for candidate images, clusters them via hierarchical K-means (~700K clusters), labels each cluster centroid with an LLM for shopping-worthiness, and publishes eligibility decisions to a distributed key-value store. The online serving path (bottom) evaluates trigger eligibility through a conjunction of Tier 0 rule-based category filtering and a Tier 1 precomputed cluster-label lookup, followed by object detection, requiring no online model inference.

4.3 Tier 1: Cluster-Level Visual Eligibility with LLM Labeling

The second tier assesses whether each image in the expanded candidate pool is visually appropriate for the shopping module. Rather than evaluating every image individually (which would be prohibitively expensive), we exploit the observation that visually similar images tend to share the same eligibility status. The offline workflow proceeds in four stages:

4.3.1 Embedding computation. Each candidate image is mapped to a fixed-dimensional vector using a visual embedding model trained on large-scale user engagement data. We selected an embedding optimized for perceptual visual similarity rather than semantic abstraction or engagement prediction [8]. This choice was deliberate: for trigger eligibility, we need images that look alike to cluster together (two outfit photos should share a label regardless of metadata differences), which differs from search applications where semantic or engagement-optimized embeddings may be preferred.

4.3.2 Clustering. Approximately 700,000 clusters are formed using hierarchical K-means on the embedding space. The hierarchical approach enables linear-time scaling with corpus size [6, 9], avoiding the quadratic complexity of standard K-means at this scale. The cluster count was chosen to balance visual homogeneity within clusters (enabling reliable label propagation) against labeling cost and storage requirements. With the candidate pool containing hundreds of millions of images, each cluster averages several hundred members, yielding roughly a 500x reduction in labeling volume compared to per-image evaluation.

4.3.3 Centroid labeling with LLM. For each cluster, the image nearest to the centroid in the embedding space is selected as the representative. A large language model is prompted to evaluate whether the representative image depicts content appropriate for a visual shopping experience. The prompt encodes the criteria for a "good" image: (a) the image should contain visible fashion or home decor items, (b) those items should be in contextual focus rather than incidental background elements, and (c) the overall composition should suggest a context where users would naturally want to explore similar products. The model returns a structured JSON response containing a binary label (good/bad).

4.3.4 Label propagation and serving. Each image in the candidate pool inherits the eligibility label of its assigned cluster centroid. The mapping from image identifier to cluster label is published to a low-latency distributed key-value store, enabling real-time lookup during serving without any online model inference. New images are incrementally assigned to the nearest existing centroid as they enter the system, and the full clustering is periodically refreshed to account for distribution shifts in new content.

Several design parameters were explored during development. The LLM model size, prompt format (free-text vs. structured JSON), output schema, embedding model, and cluster count were all treated as tunable hyperparameters. Through offline iteration comparing candidate models and prompts against human judgment on sampled images, the final configuration used a compact multimodal LLM with JSON-structured output, a visual embedding model optimized for perceptual similarity, and ~700,000 clusters, chosen to keep agreement with human judgment high while keeping labeling costs manageable.

4.4 Production Serving Path

The online serving path is intentionally simple to satisfy tight latency constraints. When a user views an image in the detail view, the system retrieves precomputed content interest scores and the T1 cluster label from the key-value store. The trigger decision is computed as a conjunction of T0 rules and the T1 label, adding <5ms to the critical path. The offline labeling pipeline runs on a weekly refresh cycle with incremental daily updates for new content, introducing a small freshness delay that is acceptable for this use case since an image's shopping-worthiness changes slowly relative to content creation rate.

This architecture trades a small amount of freshness for substantial cost and latency savings. An alternative design of running the LLM at serving time for each image would provide perfect freshness but at prohibitive cost and latency. The cluster-level approach amortizes expensive labeling across many images while maintaining high accuracy for the vast majority of content that clusters well in the embedding space.

5 Evaluation

5.1 Offline: Human Judgment Agreement

Before launching the online experiment, we conducted an offline evaluation to measure the quality of the T1 labeling pipeline. A sample of 1,000 images was randomly drawn from the candidate pool and independently labeled by human annotators as "good" or "bad" for the shopping module. The same images were processed through the full pipeline (embedding → cluster assignment → centroid label inheritance).

The evaluation measured two quantities: (1) direct label accuracy which measures the agreement between the LLM's label on the representative centroid image and a human annotator's label on the same image (91% agreement), and (2) propagation accuracy which measures the agreement between the centroid's label and the correct label for a non-centroid cluster member (∼92% agreement, with ∼8% centroid-member disagreement). The propagation error represents cases where a cluster contains visually similar but semantically distinct items near the cluster boundary which is an expected cost of cluster-level labeling that we accept given the 500x reduction in labeling volume.

Qualitative analysis of disagreement cases revealed common patterns: boundary images mixing product-oriented and non-product content, and images in inherently ambiguous categories where reasonable annotators might themselves disagree. These edge cases indicate that a meaningful share of the residual disagreement reflects genuine ambiguity in the task rather than labeler error, since some images are borderline enough that annotators themselves may disagree.

We validated the eligibility labeler against independent human judgment on 1,000 randomly sampled images (Table 1). The sample is near-balanced across classes, and the labeler reaches 91.3% agreement, well above the 56.5% majority-class baseline (Cohen's kappa = 0.82). Errors are asymmetric: the labeler rarely suppresses genuinely eligible pins (95.4% recall) but admits a larger share of ineligible ones (86.0% recall), which for a gating decision is the preferable failure mode.

Table 1: Offline agreement between LLM labels and human judgment on 1,000 randomly sampled images.

Class	Precision	Recall	Support
Eligible (Good)	89.8%	95.4%	565
Ineligible (Bad)	93.5%	86.0%	435
Overall accuracy	91.3%		1,000

5.2 Online: Iterative A/B Experimentation

The online evaluation proceeded through multiple experiment versions over several months, each building on learnings from the previous iteration. This iterative approach was essential because the trigger redesign affected a complex metric landscape spanning module-local engagement, sitewide satisfaction, and downstream shopping conversion.

Early versions tested T0 changes in isolation—expanded categories and negative filtering without the T1 precision layer. The primary learning was that category expansion alone increased module impressions but also surfaced the module on off-topic content that users found irrelevant, confirming the need for a precision layer. User engagement rates on the module actually declined in these early versions despite increased impressions, indicating that raw recall expansion without precision harms quality.

Subsequent versions introduced the T1 cluster-labeling layer alongside T0 changes. These versions validated the two-tier hypothesis: the combination of T0 recall expansion and T1 precision filtering produced stronger engagement gains than either tier alone. T0 alone improved coverage but diluted per-impression quality; T1 alone (without category expansion) maintained quality but could not reach new high-intent verticals. The combination moved both dimensions simultaneously.

Later iterations refined category exclusion thresholds based on deep-dive analysis of which content domains offered the strongest precision gains. Each iteration followed the same pattern: ship an experiment, analyze engagement patterns with domain experts, form a hypothesis about what to adjust, and test the hypothesis in the next version. This sequential, human-in-the-loop process required patience but ultimately produced a more robust configuration than any single launch could have achieved.

5.3 Final Experiment Results

The shipped configuration (T0 with expanded categories + negative filtering + T1 cluster labels + refined exclusion thresholds) was evaluated against the production baseline over four weeks. Table 2 summarizes the key metrics.

The results demonstrate that improved trigger eligibility has effects well beyond the module itself. The +11.2% module save lift indicates that users engage more with the shopping module when it appears on contextually appropriate content. The +1.0% sitewide saves lift suggests a broader halo effect: when the module is correctly triggered, it contributes positively to the overall session rather than displacing engagement from other surfaces. The sitewide save-session and revisitation lifts further confirm

Table 2: Key metrics for the shipped group vs. production baseline (four-week A/B experiment).

Metric	Level	Region	Relative Change (95% CI)	p
Saves (volume)	Module	Global	+11.2% [+6.4, +16.0]	< 0.001
Long clicks (volume)	Module	Global	+5.5% [+1.8, +9.1]	0.003
Save rate	Module	Global	+10.5% [+4.8, +16.3]	< 0.001
Trustworthy product Pin saves	Module	Global	+8.5% [+1.4, +15.6]	< 0.001
Trustworthy product Pin long clicks	Module	Global	+4.7% [+1.9, +8.4]	0.005
5-second outbound clicks	Module	UCAN	+3.5% [+0.5, +6.5]	0.023
Save sessions	Site-wide	Global	+0.43% [+0.04, +0.82]	0.029
Revisitation sessions	Site-wide	UCAN	+0.58% [+0.08, +1.09]	0.024
Saves (volume)	Site-wide	UCAN	+1.0% [+0.19, +1.82]	0.016

that users find the improved trigger experience valuable enough to increase their overall platform engagement.

Notably, module coverage—the percentage of detail page views that trigger the module—remained approximately constant at ∼15% between the baseline and the shipped treatment, despite the recall expansion in T0. This indicates that T1 precision filtering absorbed the additional candidates introduced by T0, resulting in a net redistribution of where the module appears rather than a raw increase in module frequency. The module now appears less on marginal content and more on high-quality content, yielding engagement gains at constant coverage. This is an important result: it shows that trigger quality, not trigger quantity, is the primary driver of module value.

6 Discussion

6.1 Cluster Granularity and Embedding Selection

The choice of 700,000 clusters represents a deliberate balance between label precision and operational cost. Fewer clusters (e.g., 100K) would reduce labeling cost but increase within-cluster heterogeneity, leading to more propagation errors where a cluster contains both eligible and ineligible images. More clusters (e.g., 5M) would improve homogeneity but require proportionally more LLM calls for labeling, more storage for the key-value store, and longer pipeline runtime. The 700K configuration was chosen empirically by evaluating propagation error rates at different granularities against a held-out human-labeled set.

An important practical consideration is the relationship between cluster quality and the embedding model. A high-quality embedding that places perceptually similar items close together enables coarser clustering without sacrificing label quality. Conversely, a weaker or misaligned embedding requires finer clustering to achieve the same within-cluster homogeneity. Across the embedding models we tried—including those optimized for engagement prediction, semantic similarity, and visual perceptual similarity—perceptual-similarity embeddings gave the best propagation accuracy for the eligibility task in our offline checks. This makes intuitive sense: two images that look similar to a human tend to share the same shopping-appropriateness, regardless of whether they drive similar engagement patterns or belong to the same semantic category.

6.2 When Does Cluster-Level Eligibility Work Well?

Based on our experience, cluster-level eligibility is especially attractive when: (a) per-item labels are expensive to compute (LLM calls, human annotation); (b) online latency constraints preclude real-time inference; (c) visual or semantic similarity is a useful proxy for the eligibility property; and (d) the eligibility decision changes slowly relative to the label refresh cadence. It is less attractive when eligibility depends primarily on user-specific context (personalized triggers), rapidly changing inventory (real-time availability), or fine-grained attributes that do not cluster well in available embedding spaces (e.g., price sensitivity, brand preference).

For practitioners evaluating whether cluster-level eligibility suits their use case, we recommend starting with a small-scale feasibility study: cluster a sample of items, label centroids manually or with an LLM, and measure how well the centroid label predicts the correct label for non-centroid members. If propagation accuracy is below 85%, either the embedding is misaligned with the task or the cluster granularity needs refinement. If it exceeds 90%, cluster-level labeling is likely a viable and cost-effective approach for the target application.

6.3 Iterative Experimentation as a Design Pattern

The progression through multiple experiment versions illustrates a broader lesson for production ML systems: iterative experimentation with human analysis between versions is often more productive than attempting to optimize all dimensions simultaneously [3]. Each version tested a specific hypothesis (is recall expansion sufficient alone? does the precision layer help? which exclusion thresholds produce the best trade-off?), and domain-expert analysis of failure cases informed the next version's design.

This approach was particularly valuable because the metric landscape was complex—optimizing module engagement while monitoring sitewide effects, content quality, and interactions with other product features—and no single offline metric could have predicted all the dynamics observed online [3]. We advocate for this iterative pattern as a general best practice when redesigning trigger logic for adaptive modules: plan for 3–4 experiment versions from the start, and design each version to test one clear hypothesis rather than shipping a fully-formed system on the first attempt.

7 Limitations and Future Work

The primary limitation of our system is that cluster-level labels can be incorrect for boundary items—images that are semantically distinct from their cluster centroid despite proximity in embedding space. The ~8% centroid-member disagreement is two-sided: propagation both admits some bad images and suppresses some good ones, so it is an agreement gap rather than a precision ceiling. It is reducible with finer clustering (higher cost) or item-level overrides for high-traffic or high-risk content.

A second limitation is the offline/batch nature of label updates. Newly created content receives a cluster assignment through nearest-centroid matching, but if it represents a genuinely novel visual pattern not well-represented by existing centroids, it may receive an incorrect label until the next full re-clustering. For rapidly trending content (viral images, seasonal product launches), this freshness lag could suppress valuable module impressions for hours or days.

A third limitation is that the current system treats eligibility as a static property of the image, independent of the user viewing it. Our original system design included a dynamic third tier (T2) that would incorporate session-level user intent and expected module value relative to alternative surfaces. This tier was deprioritized in favor of UX consistency: product stakeholders determined that users benefit from predictable module behavior where the same image always produces the same trigger decision, regardless of session context. As the platform develops richer intent signals and user expectations around adaptive interfaces evolve, revisiting this trade-off is a natural next step.

Future work includes two additional directions. First, replacing the binary good/bad label with a continuous eligibility score would enable threshold tuning without re-labeling, and could support personalized thresholds per user segment. Second, distilling the cluster-level labels into a lightweight neural network could eventually eliminate the key-value store dependency while preserving the quality gains from LLM-assisted labeling, and would naturally handle novel content without the freshness lag of batch re-clustering.

8 Conclusion

We presented a two-tier visual triggering system for adaptive shopping recommendations at Pinterest. By decomposing the trigger decision into a recall tier (rule-based category expansion and exclusion) and a precision tier (LLM-assisted cluster-level visual eligibility), the system improves both coverage and quality of module triggering without adding online inference cost. The precision tier draws inspiration from cluster-level generative labeling [9] but is purpose-built for the distinct requirements of binary module gating: higher accuracy demands, interaction with rule-based recall logic, and the need for iterative online validation.

A production A/B experiment demonstrated +11.2% module saves, +10.5% module save rate, and +1.0% sitewide saves—confirming that better trigger quality improves not only module-local engagement but also broad user satisfaction. Module coverage remained constant at ~15%, showing that the gains come from improved quality of trigger decisions rather than increased trigger volume. Iterative experimentation across multiple versions was essential to navigating the complex metric landscape and arriving at a robust configuration.

For the broader recommender systems community, this work demonstrates that the "when to show" decision in adaptive recommendation is a high-leverage optimization target that deserves dedicated system design. Cluster-level eligibility modeling offers a practical middle ground between brittle hand-crafted rules and expensive online learned triggers. We hope our experience—particularly the lessons around iterative experimentation, embedding selection, and the interplay between recall and precision tiers—proves useful to practitioners facing similar module gating challenges.

References

[1] Sean Bell and Kavita Bala. 2015. Learning Visual Similarity for Product Design with Convolutional Neural Networks. *ACM Transactions on Graphics* 34, 4, Article 98 (2015). doi:10.1145/2766959

[2] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine Based Neural Network for CTR Prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI '17)*. International Joint Conferences on Artificial Intelligence Organization, 1725–1731. doi:10.24963/ijcai.2017/239

[3] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, Cambridge, United Kingdom. doi:10.1017/9781108653985

[4] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. Association for Computing Machinery, New York, NY, USA, 661–670. doi:10.1145/1772690.1772758

[5] Mengze Li, Pouya Rezazadeh Kalehbasti, Sainan Chen, Minhazul Islam SK, Kshiteesh Hegde, Karina Sobhani, Kurchi Subhra Hazra, and Zhenjie Zhang. 2025. Introducing Multi-Modal Search at Pinterest: A New User Experience. In *Proceedings of the 4th End-to-End Customer Journey Workshop at KDD* (Toronto, Canada) *(KDD CJ '25)*. ACM.

[6] David Nistér and Henrik Stewénius. 2006. Scalable Recognition with a Vocabulary Tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06, Vol. 2)*. IEEE, 2161–2168. doi:10.1109/CVPR.2006.264

[7] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc.

[8] Raymond Shiau, Hao-Yu Wu, Eric Kim, Yue Li Du, Anqi Guo, Zhiyuan Zhang, Eileen Li, Kunlong Gu, Charles Rosenberg, and Andrew Zhai. 2020. Shop The Look: Building a Large Scale Visual Shopping System at Pinterest. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3203–3212. doi:10.1145/3394486.3403372

[9] Zhenjie Zhang, Sainan Chen, Kshiteesh Hegde, Mengze Li, Pouya Rezazadeh Kalehbasti, Munachim Ezema, and Karina Sobhani. 2025. Scaling Taxonomy-Free Image Labeling with Generative AI: A Multimodal Search Application. In *Proceedings of the 4th End-to-End Customer Journey Workshop at KDD* (Toronto, Canada) *(KDD CJ '25)*. ACM.

[10] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. 2018. Deep Reinforcement Learning for Page-Wise Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 95–103. doi:10.1145/3240323.3240374

Received 20 July 2026; revised 4 August 2026; accepted 25 August 2026