Introducing Multi-Modal Search at Pinterest: A New User Experience

Mengze (Irene) Li
Pinterest, Inc.
Palo Alto, CA USA
mengzeli@pinterest.com

Pouya Rezazadeh Kalehbasti
Pinterest, Inc.
Palo Alto, CA USA
pouya@pinterest.com

Sainan Chen
Pinterest, Inc.
Palo Alto, CA USA
sainanchen@pinterest.com

Minhazul Islam SK
Pinterest, Inc.
Palo Alto, CA USA
msk@pinterest.com

Kshiteesh Hegde
Pinterest, Inc.
Palo Alto, CA USA
khegde@pinterest.com

Karina Sobhani
Pinterest, Inc.
Palo Alto, CA USA
ksobhani@pinterest.com

Kurchi Subhra Hazra
Pinterest, Inc.
Palo Alto, CA USA
ksubhrahazra@pinterest.com

Zhenjie Zhang†
Pinterest, Inc.
Palo Alto, CA USA
zhenjiezhang@pinterest.com

ABSTRACT

Pinners (Pinterest users) come to Pinterest to explore their unique style and taste, and they use two major modes of search on Pinterest to do this: text search and image search. However, image and text search alone provide a disjointed set of tools to accomplish this. Pinners sometimes find it hard to express what they want through words alone. Image search also limits the user to explore topics adjacent to the reference image.

Here we introduce Multi-Modal Search: a system that accepts hybrid queries which are composed of both text and image, empowering Pinners to explore an aspect of the image more deeply (the style, the color palette, etc.) or to modify an aspect of the image (the color, the occasion, etc.). By allowing users to express their intent using both images and text, this system makes it easier for Pinners to explore their interests with greater precision. With advanced multi-modal embedding models and reranking optimizations, multi-modal search can return results that are not only relevant to the hybrid query but also engaging and diverse, thus enhancing the visual discovery experience on Pinterest. The most optimized configuration of our multi-modal search system improves the combined relevance @ 5 to the query text and image by 18% and the overall engagement by 54% compared to the baseline configuration.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence → Computer vision → Computer vision representations;
• Information systems → Information retrieval → Specialized information retrieval → Multimedia and multimodal retrieval;
• Information systems → Information retrieval → Retrieval models and ranking

KEYWORDS

Multi-modal Search, Vision Language Models (VLMs), Large Language Models (LLMs), Information Retrieval

ACM Reference format:

Mengze (Irene) 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 4th Workshop on End-End Customer Journey Optimization (KDD CJ'25). ACM, Toronto, Canada, 3 pages.* https://doi.org/XX.XXXX/XXXXXXXXXX

1 Introduction

The rise of visual discovery platforms, like Pinterest, Instagram, and TikTok, has shown visual discovery to be the ideal medium for users to find what they are looking for. However, visual discovery/search alone sometimes fails to fulfill the users' intent and get them to what they are looking for [4]. On Pinterest, for example, users often may take long journeys to get from a look/style/item they like to one that fits their taste. Through A/B testing and user studies, we have understood that providing text queries to the user which allows them to explore similar images with a shared attribute or with a changed attribute helps them find what they want more efficiently. To address this, we have introduced a new multi-modal search system at Pinterest which allows the users to find Pins (results) relevant to a text query (e.g., "vivid tones") applied to a reference image (e.g., a blue outfit) or object (e.g., a pair of black jeans). If these text queries refer to specific aspects of the reference image (e.g., color palette or fashion style), they are called "descriptors", and if they refer to

† Corresponding author

KDD CJ'25, August, 2025, Toronto, Canada
 978-1-4503-XXXX-X/18/06...
https://doi.org/XX.XXXX/XXXXXXXXXX

changes that can be applied to the reference image (e.g., changing color, occasion, or fashion style) they are called "pivots" [11]. Pivots can apply to both an entire image (aka Pin-level) or user-selected objects within an image (aka object-level). Figure 1 shows an illustration of this multi-modal search and its results.

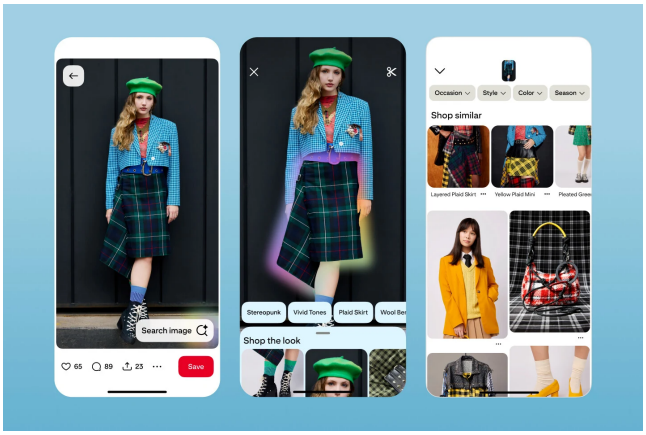

Figure 1: Illustration of the multi-modal search – image in the middle shows sample descriptors and image to the right shows sample (object-level) pivots along with results for a descriptor [2]

The multi-modal search has to find results that are, (1) both relevant to the text query and similar to the reference image/object, (2) diverse and not monotonous, and (3) engaging for the user. The first item was the most challenging, as also reported by Laenen et al. [4], Tautkute et al. [6], and Zhao et al. [12]. To address the first challenge in retrieval, we used multiple multi-modal embeddings each with different levels and modes of relevance to the text query and the reference image/object. We also included components in the reranking stage to rank at the top those candidates that were both visually similar to the reference image and complying with the queried text, allowing for results covering different variations of these two relevance dimensions. These measures in retrieval and reranking also helped with creating a heterogeneous and diverse feed for the user. Using a highly personalized ranking model also helped tackle the third challenge about providing engaging results.

2 Methodology

2.1 Retrieval

In the retrieval stage (overview shown in Figure 2), we first perform Multimodal Query Understanding on the reference image and text query pair. The information extracted from this process is used to enrich the original text query. We then use a diverse multimodal approach with three Candidate Generators (CGs) using different types of embeddings: SearchSage [5] as well as Pinterest-reproduced versions of SigLIP2 [7] and MagicLens [10]. SearchSage is trained on Pinterest Text Search Engagement data, so it excels at finding engaging results with high text relevance and contextual image relevance. The inhouse reproduced version of SigLIP2 specializes in retrieving results that resemble the reference image while being semantically relevant to the text query. Lastly, the inhouse reproduced version of MagicLens can better capture the deep interaction between the reference image and the text query, thanks to its early-fusion nature. Each CG has a complicated parameter space which makes it challenging to find a sweet spot. Therefore, we devised a Multi-Fetching mechanism that makes multiple calls to each CG with varying parameters, thus exploiting the full potential of each CG.

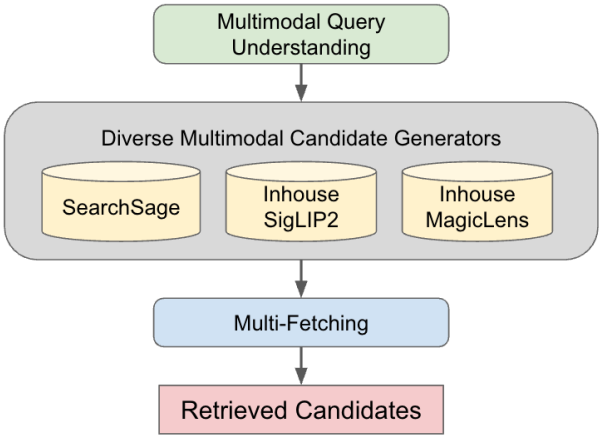

Figure 2: an overview of the retrieval stage

To enable object-level retrieval, we used Pinterest's Unified Visual Embeddings [1, 8] to find a similar image to a user-selected object, and then leveraged the pin-level pipeline to search based on this surrogate image. This enabled us to efficiently expand the pin-level search to the object level.

2.2 Ranking and Reranking

Final results from multi-modal search should be relevant to the user's text query, similar to the reference image/object, and potentially engaging given the context of the query and the reference image/object. To achieve such results, a composite reranking score (shown in the equation below) is devised that includes the cosine similarity between the embeddings of the result and reference image as well as between the result and text query. This score also includes the predicted engagement and relevance scores from Pinterest Text Search's ranking models to promote more engaging and relevant Pins given the query and the user context. Lastly, the score is multiplied by an attribute boosting factor to prioritize results with matching attributes and a CG boosting score for a balanced distribution across various CGs.

$$\begin{aligned}\text{reranking score} =& (w_1 \cdot \text{image similarity score} \\ &+ w_2 \cdot \text{text similarity score} \\ &+ w_3 \cdot \text{text search engagement score} \\ &+ w_4 \cdot \text{text search relevance score}) \\ &\cdot \sum_a (1 + w_5 \cdot \text{attribute confidence score}) \\ &\cdot \text{CG boosting score}\end{aligned}$$

After sorting based on this reranking score, the top 8 positions are filled with results that have matching attributes to the query to ensure Pinners are first presented with highly relevant content.

3 Results and Conclusion

This work introduced multi-modal search in a product used by millions of users on Pinterest, providing Pinners with a more effective tool to discover inspiration and content they enjoy. The following summarizes the gains and highlights of each mode of the multi-modal search from human evaluation and A/B experimentation, showcasing the improved efficiency and appeal compared to the other modes of search at Pinterest.

In terms of combined relevance to the reference image/object and the text query, the launched groups for pivots and descriptors had, respectively, ~14% and ~18% higher average relevance across their top 5 results compared to the baseline groups which used only a SearchSage-based candidate generator.

A/B experiments showed that the results/Pins in the multi-modal search, compared to equivalent Pins in normal text search, had 85% higher shopping engagement rate and 200% higher clickthrough rate (CTR) for descriptors, 33% higher CTR for Pin-level pivots, and 100% higher CTR for object-level pivots. Compared to image-only search [3, 9], the multi-modal search results had 2.0% higher long-term engagement propensity, 0.8% more impressions, and 0.7% higher long-impression propensity. The launched groups also showed ~54% higher overall engagement rates compared to the baseline groups.

Given the success of the current system, future work aims to expand the multi-modal search to other content verticals, customize and improve the L2 ranking model, and improve the personalization and engagement of the end-to-end multi-modal search experience.

ACKNOWLEDGMENTS

We acknowledge the help, support, and insights from our colleagues at Pinterest for the work presented in this paper: Munachin Ezema, Shirley Du, Kayoung Lee, Nikita Hudson, Ocie Henderson, Eduardo Lopez, Alejandra Estefania Garza Martinez, Helen Xu, Cindy Zhang, Anatolii Shevchenko, Rex Wu, David Xue, Eric Kim, Florian Marcu, Rajat Raina, and Sai Xiao.

REFERENCES

[1] Beal, J. et al. 2021. Billion-Scale Pretraining with Vision Transformers for Multi-Task Visual Representations. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*. (Aug. 2021), 1431–1440. DOI:https://doi.org/10.1109/WACV51458.2022.00150.

[2] Introducing new visual search features for a more personalized discovery experience | Pinterest Newsroom: *https://newsroom.pinterest.com/news/introducing-new-visual-search-features/*. Accessed: 2025-06-05.

[3] Jing, Y. et al. 2015. Visual search at pinterest. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015-August, (Aug. 2015), 1889–1898. DOI:https://doi.org/10.1145/2783258.2788621/SUPPL_FILE/P1889.MP4.

[4] Laenen, K. et al. 2018. Web search of fashion items with multimodal qerying. *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. 2018-Febuary, (Feb. 2018), 342–350. DOI:https://doi.org/10.1145/3159652.3159716.

[5] SearchSage: Learning Search Query Representations at Pinterest | by Pinterest Engineering | Pinterest Engineering Blog | Medium: *https://medium.com/pinterest-engineering/searchsage-learning-search-query-representations-at-pinterest-654f2bb887fc*. Accessed: 2025-06-05.

[6] Tautkute, I. et al. 2018. DeepStyle: Multimodal Search Engine for Fashion and Interior Design. *IEEE Access*. 7, (Jan. 2018), 84613–84628. DOI:https://doi.org/10.1109/ACCESS.2019.2923552.

[7] Tschannen, M. et al. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. (2025).

[8] Zhai, A. et al. 2019. Learning a Unified Embedding for Visual Search at Pinterest. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (Aug. 2019), 2412–2420. DOI:https://doi.org/10.1145/3292500.3330739.

[9] Zhai, A. et al. 2017. Visual discovery at Pinterest. *26th International World Wide Web Conference 2017, WWW 2017 Companion*. (2017), 515–524. DOI:https://doi.org/10.1145/3041021.3054201.

[10] Zhang, K. et al. 2024. MagicLens: Self-Supervised Image Retrieval with Open-Ended Instructions. *Proceedings of Machine Learning Research*. 235, (Mar. 2024), 59403–59420.

[11] Zhang, Z. et al. 2025. Scaling Taxonomy-Free Image Labeling with Generative AI: A Multimodal Search Application. *E2E-CJO Workshop, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2025).

[12] Zhao, Y. et al. 2022. Progressive Learning for Image Retrieval with Hybrid-Modality Queries. *SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. (Apr. 2022), 1012–1021. DOI:https://doi.org/10.1145/3477495.3532047.