PinCLIP: Large-scale Foundational Multimodal Representation at Pinterest

Josh Beal
jbeal@pinterest.com
Pinterest Inc.

Eric Kim
erickim@pinterest.com
Pinterest Inc.

Jinfeng Rao
marquisrao@pinterest.com
Pinterest Inc.

Rex Wu
rexwu@pinterest.com
Pinterest Inc.

Dmitry Kislyuk
dkislyuk@pinterest.com
Pinterest Inc.

Charles Rosenberg
crosenberg@pinterest.com
Pinterest Inc.

ABSTRACT

While multi-modal Visual Language Models (VLMs) have demonstrated significant success across various domains, the integration of the generated outputs or the underlying embedding representations directly into recommendation and retrieval systems remains a challenge, due to issues like training objective discrepancies and serving efficiency bottlenecks. This paper introduces PinCLIP, a large-scale visual representation learning approach developed to enhance retrieval and ranking models at Pinterest by leveraging VLMs to learn image-text alignment. We build upon established multi-modal models, such as CLIP [38] and SigLIP [53], and propose a novel neighbor alignment objective to optimize the similarities between two multi-modal entities (Pins) that are engaged by a similar set of users. To manage the scale of representation learning in a multi-node GPU environment, we employed several optimization techniques, including Flash Attention [13, 14, 42], funneling operations [9, 12], and activation checkpointing. The resulting representations are deployed at web-scale and integrated as features into downstream retrieval and ranking models throughout the Pinterest platform, leading to substantial improvements in relevance across all major product surfaces.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**.

KEYWORDS

PinCLIP, Visual Representation Learning, Visual Language Model, Multi-Modality, Recommendation Systems

ACM Reference Format:
Josh Beal, Eric Kim, Jinfeng Rao, Rex Wu, Dmitry Kislyuk, and Charles Rosenberg. 2026. PinCLIP: Large-scale Foundational Multimodal Representation at Pinterest. In *Proceedings of The ACM Web Conference 2026 (TheWebConf '26).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

TheWebConf '26, April 13–17, 2026, Dubai, United Arab Emirates

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

1 INTRODUCTION

Large Language Models (LLMs) have achieved great success in recent years on revolutionizing a vast number of domains starting with the introduction of GPT [39]. The introduction of the Transformer architecture [45] provides a more effective way to handle long-range dependencies in text, which led to the development of models like BERT [15], T5 [40], and GPT [39] that pioneered the use of large-scale, unsupervised pre-training to generate coherent and contextually relevant text. Subsequent breakthroughs, such as instruction tuning [10] and reinforcement learning from human feedback (RLHF) [36], have further refined these models, enabling them to follow complex instructions and produce safer, more helpful responses.

Inspired by the generative design of LLMs, early research, such as SASRec [25] or BERT4Rec [43], brings generative modeling with unidirectional or bidirectional attention through transformers [45] to sequential recommendation, though not directly leveraging the pretrained LLMs for content representations. Subsequent research, such as ZesRec [16], UniSRec [24], and others [23, 30, 41], advanced this by leveraging pre-trained LLMs to encode textual descriptions of items and user behaviors. A separate line of work [22, 37, 47] has explored using prompt-tuning [27] to directly elicit recommendations from LLMs based on user history or conversations. More recently, a trend has emerged toward pre-training large, GPT-style generative models on user interaction sequences for representation learning in recommender systems [7, 46, 48, 52].

Despite this progress, existing approaches have notable limitations. First, many models adopt LLM architectures or pre-training strategies [7, 25, 43, 46, 52, 56] without directly harnessing the rich semantic representations from pre-trained foundation models. Second, the primary focus has remained on textual data [16, 23, 24, 30, 41], with other modalities such as images and video being largely underexplored. Finally, most studies are conducted on small-scale academic datasets, providing insufficient insight into the challenges of deploying these models scalably in real-world production environments.

Overall, we believe that there are several fundamental challenges that impede the integration of LLMs into production recommender systems (RecSys). Primarily, the optimization objectives differ between the two paradigms. Modern LLMs like GPT [39] are generative models trained to predict the next token, whereas RecSys models are typically trained for discriminative tasks – such as predicting user engagement with a video – utilizing cross-entropy loss functions. Secondly, the operational requirements of RecSys necessitate high throughput and low latency to serve live traffic.

The substantial model sizes of LLMs, which can contain billions of parameters, pose a practical challenge for deployment in production RecSys due to their considerable computational cost and inherent latency. Finally, LLMs and large vision models (VLMs) are trained on multi-modal data, where inputs are tokenized into granular units like text tokens or image patches. In contrast, RecSys models leverage manually engineered features and learn from sparse implicit feedback derived from a potentially vast corpus of items. These critical architectural and data-centric differences present substantial technical hurdles for the seamless integration of LLMs into RecSys.

This paper introduces **PinCLIP**, a large-scale visual representation learning approach developed to enhance retrieval and ranking models at Pinterest. We build upon established multi-modal models, such as CLIP [38] and SigLIP [53], and propose a novel neighbor alignment objective to optimize the similarities between two images that are engaged by a similar set of users. To manage the scale of representation learning, we employed several optimization techniques, including Flash Attention [13, 14, 42], funneling operations [9, 12], and activation checkpointing to enable multi-node training. The resulting representations are integrated into dozens of machine learning production models, significantly enhancing user engagement across different surfaces in Pinterest.

The key contributions of this work are as follows:

- We introduce a novel approach that leverages pretrained VLMs to learn visual representations from a massive, real-world product with hundreds of millions of users.
- We present a practical recipe on how to build an effective multi-modality content signal on a large-scale image dataset, including our best practices on data selection, model architecture, learning objectives, efficiency enhancements and full productionization.
- We validate the general effectiveness of our proposed visual representations not only by strong offline metric gains, also by numerous online integrations to various surfaces in Pinterest, including Homefeed, Search and Related Pins.
- We also show that our learned multi-modality content representations are highly effective at promoting fresh content on the platform with up to 15% more user actions, which is a crucial step toward solving the longstanding cold-start problem in recommender systems.

2 RELATED WORK

2.1 Visual Language Models

The evolution of visual-language models (VLMs) began with a foundational shift from CNNs to powerful Transformer-based vision backbones like the Vision Transformer (ViT) [17], which treated image patches as sequential tokens, and the more efficient Swin Transformer [33], which introduced a hierarchical, windowed self-attention mechanism. These architectures paved the way for large-scale pre-training strategies, exemplified by OpenAI's CLIP [38], which pioneered contrastive learning on massive web-scale image-text pairs to achieve remarkable zero-shot generalization. Subsequent models like BLIP [28] refined this approach by introducing image-to-text generation tasks, and SigLIP [53] simplified contrastive learning with sigmoid loss to achieve tremendous efficiency gains. The current frontier is defined by scaling, with public models such as InternVL [8, 57] and Alibaba's Qwen-VL [3], and closed models like Gemini 2.5 Pro [11], pushing parameter counts into the billions. These state-of-the-art systems demonstrate that increasing model and data scale continues to yield significant gains, enabling more sophisticated and fine-grained multimodal reasoning capabilities.

2.2 LLMs for Recommender Systems

Drawing inspiration from the generative capabilities of LLMs, initial studies in sequential recommendation adopted Transformer architectures. Models like SASRec [25], BERT4Rec [43], and S^3-Rec [56] utilized generative modeling with either unidirectional or bidirectional attention, although they did not use pre-trained LLMs for content understanding. A later wave of research, including ZesRec [16], UniSRec [24] and others [23, 30, 41], advanced the field by employing pre-trained LLMs to encode the textual features of items and user interactions. Concurrently, another research direction explored prompt-tuning LLMs to directly generate recommendations from user histories or dialogues, as seen in works by [22, 37, 47]. More recently, the field has shifted towards pre-training large, GPT-style generative models directly on sequences of user behavior to learn powerful representations for recommendation tasks [7, 46, 48, 52]. Fu et al. [19] utilize multi-modality models as item encoders for sequential recommendation and introduce a decoupled PEFT mechanism to boost efficiency. Their approach still requires multi-modality model inferences for each item in the user sequence, which could lead to significant scalability challenges when scaling to a production environment with hundreds of millions of users and items.

A concurrent work from Giahi et al. [20] also learns multi-modal embeddings using a CLIP-style objective by aligning object-level image crops with LLM-generated text summaries. Our approach differs in several critical aspects. First, unlike their method which models vision and text through separate encoders, our model employs a cross-modal fusion architecture to learn a richer, unified semantic representation. Second, we introduce a novel neighbor alignment task that is critical for learning effective representations. Consequently, our approach produces a holistic image embedding not explicitly defined in their object-centric method, demonstrating strong performance on both retrieval and ranking tasks, whereas their evaluation is focused primarily on retrieval.

3 METHOD

Pinterest, as a visual-centric platform, is primarily used by its users to discover visual inspiration across a wide spectrum of topics. To enhance the functionality of the platform, our objective is to construct a general and effective multimodal representation learning system. This system is designed to benefit numerous downstream machine learning models in production, spanning various surface areas such as Homefeed, Related Pins, Search, and Ads. To achieve this goal, we propose and develop two distinct model architectures leveraging a contrastive learning framework:

- **Image-Text Model**: This architecture utilizes image-text contrastive learning exclusively, following the methodology pioneered by OpenAI's CLIP [38]. It is trained on pairs

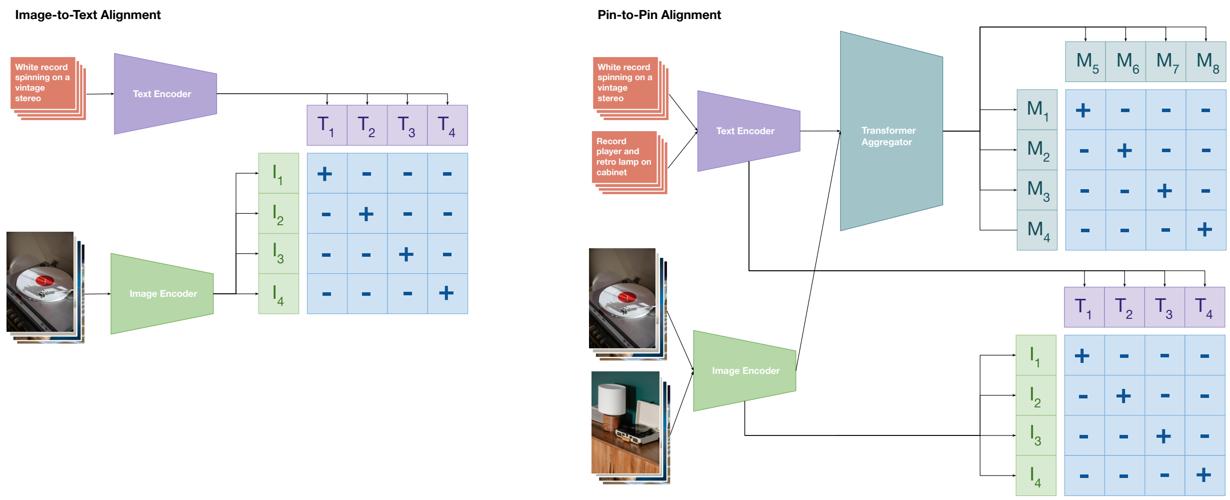

Figure 1: Side-by-side comparison of image-text and fusion objectives. The image-text model (left) uses a contrastive learning strategy inspired by CLIP [38] and SigLIP [53], whereas the fusion model (right) incorporates a new objective that aligns multimodal Pin representations. This new objective leads to stronger offline and online improvements for the fusion model.

comprising Pinterest images and associated textual metadata, which includes titles, detailed descriptions, search keywords, and internal annotations.

- **Fusion Model**: This architecture employs a more sophisticated multimodal contrastive learning objective. It explicitly incorporates the fusion of image and text information to generate a unified representation, followed by an alignment step to refine the overall Pin representation.

3.1 Image-Text Model Architecture

In this section, we describe the image-text model in detail. As shown on the left in Figure 1, it consists of an image encoder $E_{\text{img}}(\cdot)$ and a text encoder $E_{\text{txt}}(\cdot)$, where the same text encoder is used to process different text sources. We use multiple contrastive loss terms to align: (1) the query image and the descriptive text and (2) the query image and the keyword text. In particular, the descriptive text consists of the short Pin title, the detailed Pin description, or the synthetic caption (generated by an open-source VLM [29, 31]), selected in order of availability. Likewise, the keyword text consists of the engaged search keywords or the extracted Pin keywords, selected in order of availability. We found that it was more effective to coalesce the text sources as described above, rather than concatenate all of the available text into a single target string. Several works have observed that CLIP text encoders exhibit a relatively short effective context length [55], and highly detailed image descriptions can reduce the performance of contrastive learning [32].

As in previous work, we have multiple choices for the image and text encoders. We consider two in-house pretrained image encoders [4] of varying sizes, which significantly outperform open-source vision models for this application. These models consist of a convolutional neural network stem and a Funnel Vision Transformer trunk [12, 17]. We consider two open-source text encoders of varying sizes, specifically the SigLIP multilingual text encoders [53].

We adopt the sigmoid loss from SigLIP, which is a simple pairwise loss that outperforms the traditional softmax loss used in previous work [38]. This objective yields better downstream performance, especially at smaller batch sizes, while also improving training efficiency. By processing each image-text pair independently, it avoids the heavy communication between nodes that is required to compute the global normalization factor. We describe the efficient "chunked" implementation below, which further reduces the significant communication costs of multi-node training.

Consider a mini-batch $\mathcal{B} = ((I_1, T_1), \ldots, (I_n, T_n))$ of image-text pairs, and let $\mathcal{B}' = ((I'_1, T'_1), \ldots, (I'_n, T'_n))$ be the encoded representations where $I'_j = E_{\text{img}}(I_j)$ and $T'_j = E_{\text{txt}}(T_j)$ for $j \in \{1, \ldots, n\}$. For a set of devices of size D, we let $b = \frac{|\mathcal{B}|}{D}$ be the per-device batch size. We define the per-instance loss as:

$$\mathcal{L}_{ij} = \log(1 + e^{\mathbf{z}_{ij}(-t\mathbf{x}_i \cdot \mathbf{y}_j + c)})$$

where $\mathbf{x}_i = \frac{I'_i}{\|I'_i\|_2}$ and $\mathbf{y}_j = \frac{T'_j}{\|T'_j\|_2}$ are the normalized image/text representations, and $\mathbf{z}_{ij}$ is the label for a given image-text pair, equaling 1 if they originate from the same Pin and -1 otherwise, t is the temperature, and c is the bias. To improve convergence, we set $t = \log 10$ and $c = -10$ at initialization.

We define the chunked sigmoid loss as follows:

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|} \sum_{d_i=1}^{D} \sum_{d_j=1}^{D} \sum_{i=bd_i}^{b(d_i+1)} \sum_{j=bd_j}^{b(d_j+1)} \mathcal{L}_{ij}.$$

This is equivalent to the following simplified loss calculation, which materializes the full embedding similarity matrix:

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \mathcal{L}_{ij}.$$

3.2 Fusion Model Architecture

The fusion model introduces the following key modifications to the image-text model. First, we introduce a new objective that aligns the representations of query Pins and target Pins. These Pin pairs are sampled from a large-scale (pruned) Pin-Board graph, where an edge exists between a Pin and a Board if a user has saved the Pin to a given Board, which is a collection of Pins on the same topic. This manual curation signal yields an extremely valuable bipartite graph, which has been widely used at Pinterest for efficient candidate generators [18] and engagement-based embeddings [2, 18, 49].

The fusion encoder $E_{\mathrm{fsn}}(\cdot)$ computes the image and text representations of the Pin, and applies a Transformer aggregator, consisting of 2 layers, to the unpooled visual and textual tokens, yielding a multimodal embedding $M'_j = E_{\mathrm{fsn}}(I_j, T_j)$. For this task, we use the descriptive text associated with the Pin. We adapt the sigmoid loss to enforce the similarity of the paired multimodal representations. As above, we normalize the resulting multimodal embedding. We let $\mathbf{u}_i = \frac{M'_i}{\|M'_i\|_2}$ and $\mathbf{v}_j = \frac{M'_{j+n/2}}{\|M'_{j+n/2}\|_2}$ are the normalized multimodal representations for $i, j \in [\frac{n}{2}]$ and $\mathbf{w}_{ij}$ is the label for a given pair of Pins, equaling 1 for neighbor Pins and -1 otherwise, t is the temperature, and c is the bias. We extend the pairwise loss from [53] and define the per-instance loss as:

$$\mathcal{L}_{ij} = \log(1 + e^{\mathbf{w}_{ij}(-t\mathbf{u}_i \cdot \mathbf{v}_j + c)})$$

3.3 Training Data

The training dataset consists of two main formats: image-text pairs and Pin neighbor pairs. The first format supports the traditional image-text alignment objective. The second format is introduced to support the multimodal fusion objective.

3.3.1 Image-Text Pairs. Each image is paired with text captions that describe the image. The text captions come from a variety of sources, such as user-provided (e.g., Pin titles and descriptions) or generated by an image captioning model. We choose the highest quality text source for each Pin. There are 889,767,355 unique images in the PinText dataset. To improve the quality of the image-text pairs, we calculated image-text alignment scores (via the CLIPScore [38] method using OpenAI's CLIP-L as the reference model) and filtered out all pairs whose alignment score was below a tuned threshold.

3.3.2 Neighbor Pairs. For each query Pin, we sample $N = 5$ similar Pins to produce N query-positive pairs. The intent is that each query Pin and positive Pin should be semantically similar to each other. The sampling is done by performing random walks over the Pin-Board graph, starting at the query node, aggregating the visited nodes, and choosing the top N visited nodes. For context, the Pin-Board graph is a graph data structure where each node is a Pin or a Board, and an (undirected) edge connects a Pin to a Board if a Pin is present in a Board. For efficiency, we first compute the top $K = 50$ pairs for each query offline, and then sample N pairs from the cached results. We observed a small benefit to weighted sampling by the visit count vs. uniform random sampling..

3.4 Training Efficiency

The vision encoder design is a critical factor in improving the training efficiency. Adopting a simple Vision Transformer backbone dramatically increases the GPU resource requirements (by a factor of 2-4x) to achieve the same level of performance. Furthermore, we observed benefits from partial freezing of the vision encoder. This approach resulted in no degradation of the model quality in our experiments. In fact, freezing several Transformer layers was found to improve the downstream performance, similar to prior findings on locked image-tuning (LiT) [54]. We attribute this result to the high quality of the pretrained in-house image encoder, which is trained on a multi-label classification task [4]. To further improve the training efficiency, we adopted FlashAttention-2 [13] and activation checkpointing techniques, which improved training throughput by 15% and reduced GPU memory usage by 32%.

3.5 Signal Productionization

To reduce serving costs and ease signal adoption for downstream consumers, we explored reducing the embedding dimensionality via Matryoshka Representation Learning (MRL) [26]. As the model was trained with the MRL loss for prefixes 64d (64 dimension), 128d, and 256d (full), we publish reduced-dimensionality embeddings for downstream consumers to ingest, since ingesting the full 256d embeddings could be infeasible for cost and resource reasons. We found that the 64d prefix embedding offers an acceptable tradeoff between representation quality and serving costs. To ensure that the embeddings can be easily utilized for downstream retrieval systems like approximate nearest neighbor search via inner-product distance, we perform L2 normalization of the prefix embedding.

We additionally found that 8-bit affine quantization resulted in negligible impact on retrieval quality, yielding further cost savings.

4 EXPERIMENTS

4.1 Evaluation Setup

We design our offline evaluation suite to reflect the real-world performance of the model in the production systems at Pinterest. We consider five key evaluation tasks: PinText image-text retrieval, Related Pins image-image retrieval, Related Pins multimodal retrieval, Search text-image retrieval and Search multimodal retrieval. For the multimodal retrieval tasks, we use the fusion embedding as the Pin representation, whereas the image embedding is used for the other surface evaluations.

We sample 30k pairs from the PinText training dataset to form the PinText evaluation dataset. For Related Pins and Search, we sample 80k engaged pairs, where a user saved the target Pin for a given text search query or query Pin. We include a 1.5M distractor set for the surface evaluations. For a proper evaluation, we ensure that the evaluation sets are disjoint from the training sets.

We compute the Recall@K metric for each evaluation task. Specifically, we evaluate whether the proposed method can retrieve the correct positive embedding among a set of random negative embeddings, given the query embedding. Given a set of query embeddings Q, positive embeddings P, and negative embeddings N, the Recall@K metric is defined below, where $|Q|$ denotes the number of query embeddings and the similarity metric is $s(n_1, n2) = n_1^T n_2$.

$$\mathrm{Recall}_K(Q, P, N) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathbf{1}\{|\{\, n \in N \mid s(Q_i, n) \geq s(Q_i, P_i) \,\}| < K\}$$

Image Encoder	Text Encoder	R@1	R@5	R@10
Image-to-Text Retrieval				
Hybrid-ViT-B (130M)	Transformer-B (278M)	56.2	71.1	75.7
Hybrid-ViT-B (130M)	Transformer-L (700M)	60.3	73.9	78.3
Hybrid-ViT-g (922M)	Transformer-B (278M)	67.2	79.3	82.7
Hybrid-ViT-g (922M)	Transformer-L (700M)	70.3	81.8	84.9

Table 1: Assessment of the impact of model scaling per modality on PinText image-to-text retrieval performance.

4.2 Training Procedure

We train with a global batch size of 32,768 for the image-text alignment task and 4,096 for the neighbor alignment task using 64 H100 GPUs. We train for 150,000 steps, processing ≈6B images in total. We use a warmup phase of 6,000 steps, followed by cosine decay of the learning rate to zero. The base learning rate is 2×10^{-4}, and the optimizer is Lion [6]. We found it beneficial to employ a 10x lower learning rate for the text encoder. We used a weight decay value of 5×10^{-4} for the image encoder and fusion aggregator and a weight decay value of 5×10^{-1} for the text encoder based on grid search.

4.3 Main Results

The main results are presented in Figure 2. We find that the PinCLIP Fusion model significantly outperforms the PinCLIP image-text model on the surface retrieval tasks (+39% Search, +44% Related Pins) due to the introduction of (1) the cross-modal fusion architecture and (2) the neighbor alignment objective. Our method significantly outperforms prior internal content-only embedding models, such as OmniSearchSage [1] and Unified Embedding [51], and state-of-the-art public multimodal embedding models [3, 5, 38, 44, 53].

4.4 Ablation Studies

For these experiments, we train with a shorter training schedule consisting of 30k steps (≈1B examples seen) using the PinCLIP Fusion model architecture, unless otherwise stated.

4.4.1 Impact of Model Size. To understand the impact of model size on the task performance, we trained models at four scales.

We train with two different in-house image encoders, each of which is pretrained on a multi-label classification task [4]. "Hybrid-ViT-B" (130M) is a smaller image encoder consisting of a ResNet-9 [21] stem and a "Funnel-ViT-Base-6x3" trunk. Specifically, the Vision Transformer consists of 3 modules with 2 intermediate funneling operations. The hidden dimension is 768, the MLP dimension is 3072, the layers per module is 6, and the number of heads is 12. "Hybrid-ViT-g" (922M) is a larger image encoder consisting of an InceptionNeXt-Tiny [50] stem and a "Funnel-ViT-giant-12x3" trunk. In this case, the hidden dimension is 1408, the MLP dimension is 6144, the layers per module is 12, and the number of heads is 16.

We train with two different multilingual pretrained text encoders from SigLIP [53]. "Transformer-B" (278M) is a smaller text encoder consisting of 12 layers with width of 768 and 12 heads. "Transformer-L" (700M) is a larger text encoder consisting of 27 layers with width of 1152 and 16 heads. The vocabulary size is 250k for each model.

The results are presented in Table 1. We observe consistent gains from increased model capacity for each modality.

(a) PinText

Method Name	R@1
Image-to-Text Retrieval	
CLIP-B [38]	29.4
CLIP-L [38]	35.7
mSigLIP-So [53]	58.6
PEcore-G [5]	60.5
SigLIP2-g [44]	62.0
MetaCLIP2-g [35]	63.8
Qwen-ViT [3]	60.9
PinCLIP Image-Text	**76.7**
PinCLIP Fusion	<u>76.4</u>

(b) Search

Method Name	R@10
Text-to-Image Retrieval	
CLIP-B [38]	9.3
CLIP-L [38]	12.4
mSigLIP-So [53]	24.3
PEcore-G [5]	23.6
SigLIP2-g [44]	26.7
MetaCLIP2-g [35]	27.4
Qwen-ViT [3]	29.7
PinCLIP Image-Text	32.0
PinCLIP Fusion	**41.6**
Multimodal Retrieval	
OmniSearchSage [1]	34.9
Qwen-ViT [3]	39.7
PinCLIP Fusion	**44.5**

(c) Related Pins

Method Name	R@10
Image-to-Image Retrieval	
CLIP-B [38]	19.7
CLIP-L [38]	22.7
mSigLIP-So [53]	35.1
PEcore-G [5]	35.3
SigLIP2-g [44]	35.7
MetaCLIP2-g [35]	33.8
Qwen-ViT [3]	29.1
Unified Embedding [51]	31.4
PinCLIP Image-Text	40.2
PinCLIP Fusion	**48.2**
Multimodal Retrieval	
Qwen-ViT [3]	38.5
PinCLIP Fusion	**58.2**

Figure 2: Retrieval results (Recall@K) for the key evaluation tasks of PinText, Related Pins, and Search.

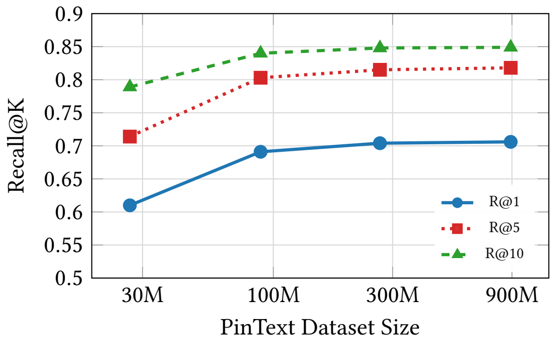

Figure 3: Assessment of the impact of dataset scaling on PinText image-to-text retrieval performance.

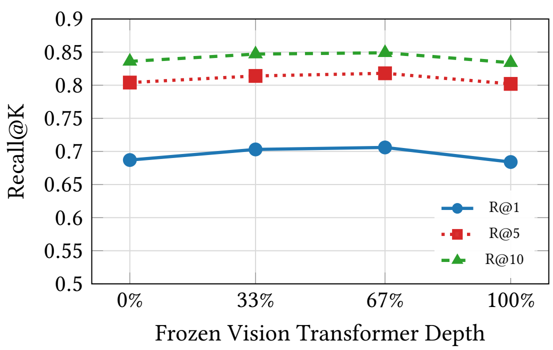

Figure 4: Assessment of the impact of freezing image encoder layers on PinText image-to-text retrieval performance.

4.4.2 Impact of Dataset Size. We study the impact of dataset size, varying the dataset sampling ratio from 3% to 100% and observing the impact on image-text retrieval performance. In these experiments, the number of samples seen is held constant, with approximately 1B samples seen in these experiments. The results are presented in Figure 3, showing a clear benefit to larger datasets.

4.4.3 Impact of Freezing Layers. In the contrastive tuning paradigm explored by LiT [54], a strong, pretrained image encoder is used as the image tower. Over the course of training, the image tower and text tower weights may be locked or unlocked. While LiT [54] proposes locking all of the image tower weights, we find that unlocking several of the final layers yields better performance. We study the optimal number of image encoder layers to unlock in Figure 4. We find that unlocking the final 12 layers of the image encoder yields the best downstream performance while also significantly improving the training efficiency. Consistent with the findings in LiT [54], we saw no benefit to locking layers of the text tower. In Table 2, we report the impact on GPU memory usage and training throughput for a single-node training benchmark setup.

Locked Layers	GPU Memory (GB)	Throughput (img/s)
0	69.37	1223
12	58.19 (-16.1%)	1418 (+15.9%)
24	48.67 (-29.8%)	1572 (+28.5%)
36	41.38 (-40.3%)	1658 (+35.5%)

Table 2: Assessment of the impact of freezing image encoder layers on 1-node training efficiency (memory/throughput).

Online Metric	Homefeed	Related Pins	Search
Sitewide Repins	**+0.20%**	**+0.37%**	N/A
Surface Repins	**+0.91%**	**+1.84%**	**+0.96%**
Surface Fresh Repins	**+5.02%**	**+14.15%**	**+15.35%**

Table 3: Online A/B tests by adding PinCLIP as a feature into the ranking models in Homefeed, Related Pins and Search.

Related Pins		**Search**
Sitewide Repins	Sitewide Fresh Repins	Fulfillment Rate
+0.36%	**+2.3%**	**+0.34%**

Table 4: Online A/B tests by using PinCLIP-based candidate generator in Related Pins and Search.

4.5 Online A/B Testing

To evaluate the utility of our proposed multi-modal embeddings, we perform online A/B experiments across the major surfaces at Pinterest, including Homefeed, Related Pins and Search.

4.5.1 Ranking Models. We added PinCLIP as a new feature into the ranking models in different surfaces. We ran A/B online experiments over more than 2 weeks of data to collect the online impact. The control group is the production ranking model without the PinCLIP feature, and the treatment group is the production model with PinCLIP. For the metrics, we measure surface-level and sitewide repins, which is the most common user action type at Pinterest. Meanwhile, we also measure the number of user repins on fresh content on each surface, where fresh content is defined as new content with age less than 28 days.

The online results are shown in Table 3. We see significant gains in surface repins for all three main surfaces on Pinterest, which also led to significant site-wide repin gains on Homefeed and Related Pins. The sitewide repins for Search is not significant as the surface contributes fewer repins, relative to the other surfaces. We also see PinCLIP is extremely effective on distributing fresh content, leading to about 15% fresh repin gains on Related Pins and Search.

4.5.2 Candidate Generators. We study how to leverage PinCLIP as a new candidate generator to retrieve candidates on Search and Related Pins. On Search, we used the text embedding produced by the PinCLIP model to encode the search query, and used the fusion embedding from the PinCLIP model to encode the image corpus. On Related Pins, we used the fusion embedding to encode both the query pin and the image corpus. For all candidate generator experiments, we used the approximate nearest neighbor (HNSW) [34]

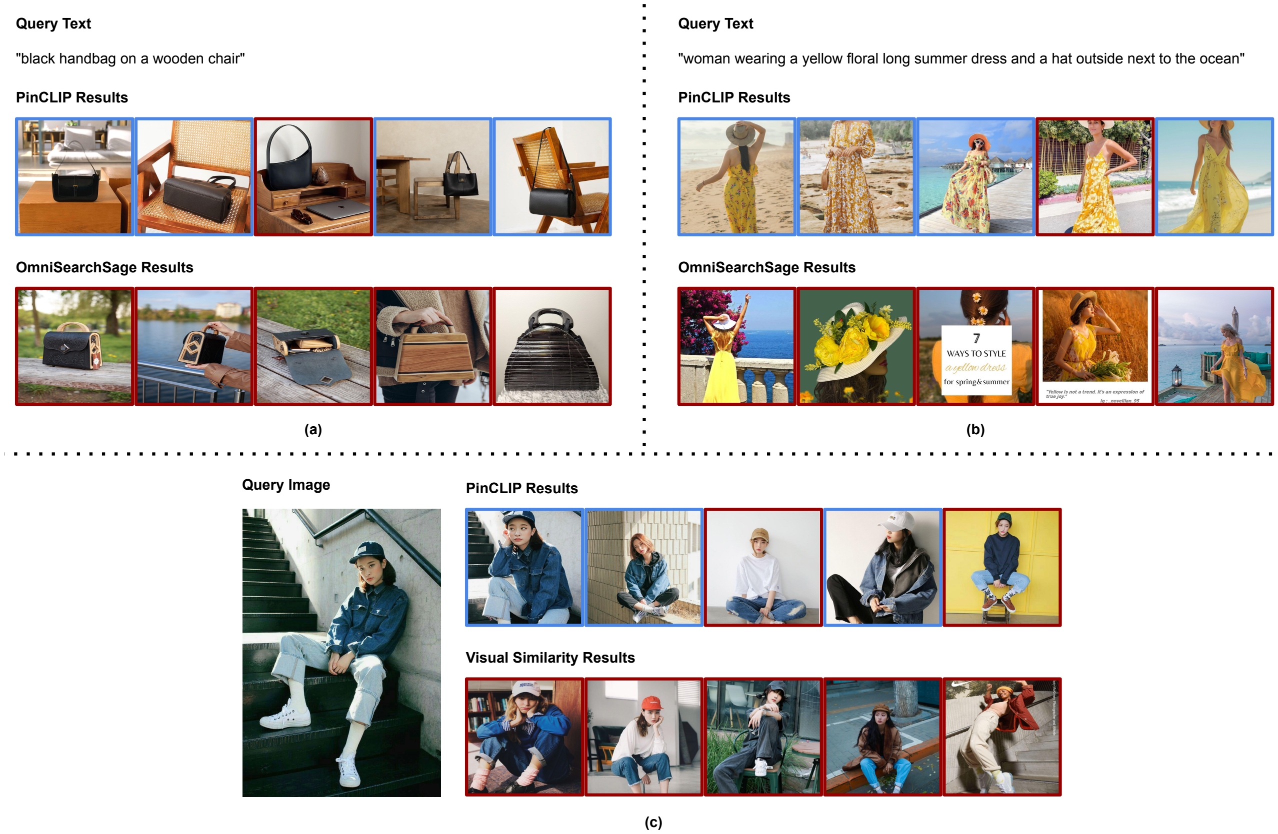

Figure 5: Qualitative comparison of PinCLIP and OmniSearchSage [1] retrieved candidates on production-scale search corpus. Relevant candidates are highlighted in blue, whereas irrelevant candidates are highlighted in red. PinCLIP produces more semantically similar results compared to competitive baselines for both using text queries or image as inputs.

algorithm to retrieve the top k candidates based on the similarities of the dot product between the candidate embedding and the context embedding (search query or query pin).

As we can see in Table 4, the addition of the PinCLIP candidate generator in Related Pins led to a significant improvement in the number of repins throughout the site and the number of fresh repins throughout the site. In Search, the fulfillment rate is defined as the percentage of search sessions with positive user engagement actions, such as repins, shares, long-clicks, etc. We also observed a significant gain in the search fulfillment rate. These experiments have demonstrated the general effectiveness of PinCLIP on both early-stage retrieval and late-stage ranking models.

4.6 Qualitative Study

In Figure 5, we compare the retrieved results between the OmniSearchSage [1] and our PinCLIP fusion model, when using both text queries or image as input. As we can see, in Figure 5(a), PinCLIP successfully preserves object attributes (color, material) along with object-level relationships ("bag *on* a chair"). In (b) PinCLIP successfully handles long text queries with multiple details. In (c), we compare PinCLIP results against our existing image embedding retrieval system. While both embeddings perform well on visual queries, we generally find that PinCLIP can yield better results. These qualitative studies further validate the general effectiveness of our proposed approach.

5 CONCLUSION

This paper introduces PinCLIP, a large-scale visual representation learning approach developed by Pinterest to enhance its retrieval and ranking recommendation models. The core innovation lies in enhancing established multi-modal models like CLIP [38] and SigLIP [53] through a novel neighbor alignment objective. This objective leverages valuable implicit user feedback from Pinterest's Pin-Board graph to ensure that Pins engaged by similar users have closer representations. The paper also provides a practical solution for scalably training and serving these multi-modal representations within a production recommendation system. Online experiments have validated the superior effectiveness of PinCLIP embeddings, showing significant improvements in engagement metrics across multiple surfaces for both retrieval and ranking models, most notably an increase of up to 15% in fresh repins (saves for recently created content), which is crucial for solving the cold-start problem in recommender systems.

REFERENCES

[1] Prabhat Agarwal, Minhazul Islam Sk, Nikil Pancha, Kurchi Subhra Hazra, Jiajing Xu, and Chuck Rosenberg. 2024. OmniSearchSage: Multi-Task Multi-Entity Embeddings for Pinterest Search. In *Companion Proceedings of the ACM Web Conference 2024*. 121–130.

[2] Anirudhan Badrinath, Alex Yang, Kousik Rajesh, Prabhat Agarwal, Jaewon Yang, Haoyu Chen, Jiajing Xu, and Charles Rosenberg. 2025. OmniSage: Large scale, multi-entity heterogeneous graph representation learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 4261–4272.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

[4] Josh Beal, Hao-Yu Wu, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. 2022. Billion-scale pretraining with vision transformers for multi-task visual representations. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 564–573.

[5] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. 2025. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181* (2025).

[6] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. 2023. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems* 36 (2023), 49205–49233.

[7] Xiangyi Chen, Kousik Rajesh, Matthew Lawhon, Zelun Wang, Hanyu Li, Haomiao Li, Saurabh Vishwas Joshi, Pong Eksombatchai, Jaewon Yang, Yi-Ping Hsu, et al. 2025. PinFM: Foundation Model for User Activity Sequences at a Billion-scale Visual Discovery Platform. *arXiv preprint arXiv:2507.12704* (2025).

[8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271* (2024).

[9] DongHyun Choi, Lucas Spangher, Chris Hidey, Peter Grabowski, and Ramy Eskander. 2025. Revisiting Funnel Transformers for Modern LLM Architectures with Comprehensive Ablations in Training and Inference Configurations. *arXiv preprint arXiv:2504.02877* (2025).

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53.

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[12] Zihang Dai, Guokun Lai, Yiming Yang, and Quoc Le. 2020. Funnel-transformer: Filtering out sequential redundancy for efficient language processing. *Advances in neural information processing systems* 33 (2020), 4271–4282.

[13] Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691* (2023).

[14] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems* 35 (2022), 16344–16359.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[16] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-shot recommender systems. *arXiv preprint arXiv:2105.08318* (2021).

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[18] Chantat Eksombatchai, Pranav Jindal, Jerry Zitao Liu, Yuchen Liu, Rahul Sharma, Charles Sugnet, Mark Ulrich, and Jure Leskovec. 2018. Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *Proceedings of the 2018 world wide web conference*. 1775–1784.

[19] Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Jie Wang, and Joemon M Jose. 2024. IISAN: Efficiently adapting multimodal representation for sequential recommendation with decoupled PEFT. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 687–697.

[20] Ramin Giahi, Kehui Yao, Sriram Kollipara, Kai Zhao, Vahid Mirjalili, Jianpeng Xu, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. 2025. VL-CLIP: Enhancing Multimodal Recommendations via Visual Grounding and LLM-Augmented CLIP Embeddings. *arXiv preprint arXiv:2507.17080* (2025).

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[22] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 720–730.

[23] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952* (2024).

[24] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 585–593.

[25] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[26] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems* 35 (2022), 30233–30249.

[27] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[30] Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. 2023. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1258–1267.

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 26296–26306.

[32] Yanqing Liu, Xianhang Li, Zeyu Wang, Bingchen Zhao, and Cihang Xie. 2024. Clips: An enhanced clip framework for learning with synthetic captions. *arXiv preprint arXiv:2411.16828* (2024).

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[34] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836.

[35] Jean de Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding Multilingual Multimodal LLMs With Cultural Knowledge. *arXiv preprint arXiv:2508.07414* (2025).

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[37] Gustavo Penha and Claudia Hauff. 2020. What does bert know about books, movies and music? probing bert for conversational recommendation. In *Proceedings of the 14th ACM conference on recommender systems*. 388–397.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[39] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[41] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM web conference 2024*. 3464–3475.

[42] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems* 37 (2024), 68658–68685.

[43] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international*

conference on information and knowledge management. 1441–1450.
[44] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786* (2025).
[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[46] Chunqi Wang, Bingchao Wu, Zheng Chen, Lei Shen, Bing Wang, and Xiaoyi Zeng. 2025. Scaling transformers for discriminative recommendation via generative pretraining. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 2893–2903.
[47] Bowen Yang, Cong Han, Yu Li, Lei Zuo, and Zhou Yu. 2021. Improving conversational recommendation systems' quality with context-aware item meta information. *arXiv preprint arXiv:2112.08140* (2021).
[48] Chao Yi, Dian Chen, Gaoyang Guo, Jiakai Tang, Jian Wu, Jing Yu, Sunhao Dai, Wen Chen, Wenjun Yang, Yuning Jiang, et al. 2025. RecGPT Technical Report. *arXiv preprint arXiv:2507.22879* (2025).
[49] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 974–983.
[50] Weihao Yu, Pan Zhou, Shuicheng Yan, and Xinchao Wang. 2024. Inceptionnext: When inception meets convnext. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 5672–5683.
[51] Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, and Charles Rosenberg. 2019. Learning a unified embedding for visual search at pinterest. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2412–2420.
[52] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, et al. 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152* (2024).
[53] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.
[54] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18123–18133.
[55] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*. Springer, 310–325.
[56] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1893–1902.
[57] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025).