Pinecone

Building remarkable multimodal search applications with Pinecone and AWS

Available in AWS Marketplace

Pinecone

Table of contents

Why use multimodal search?

The artificial intelligence (AI) landscape is currently dominated by generative AI, capturing much of the spotlight. However, the market is ripe with opportunities for practical applications that go beyond content generation. Multimodal search exemplifies this potential, offering the ability to process and understand diverse data types simultaneously. This capability paves the way for more comprehensive and contextually rich information retrieval and analysis, opening new avenues for innovation and value creation in the AI industry.

What is multimodal search?

Multimodal search makes it possible to use more than just text in queries, including images, audio, and videos. It can represent high-level, semantic relationships and connections across data types and formats (data modalities). The concept of embeddings for a single data type is not new. Clustering convolutional neural network (CNN) image embeddings to find visually similar images has been used for years. But advances in tech have revolutionized multimodal.

With multimodal vector search, for example, results are based on the similarity of the content and representations in the text, image, and more. How does multimodal vector search do this? It encodes and relates multiple different modalities into a shared embedding space (see Figure 1).

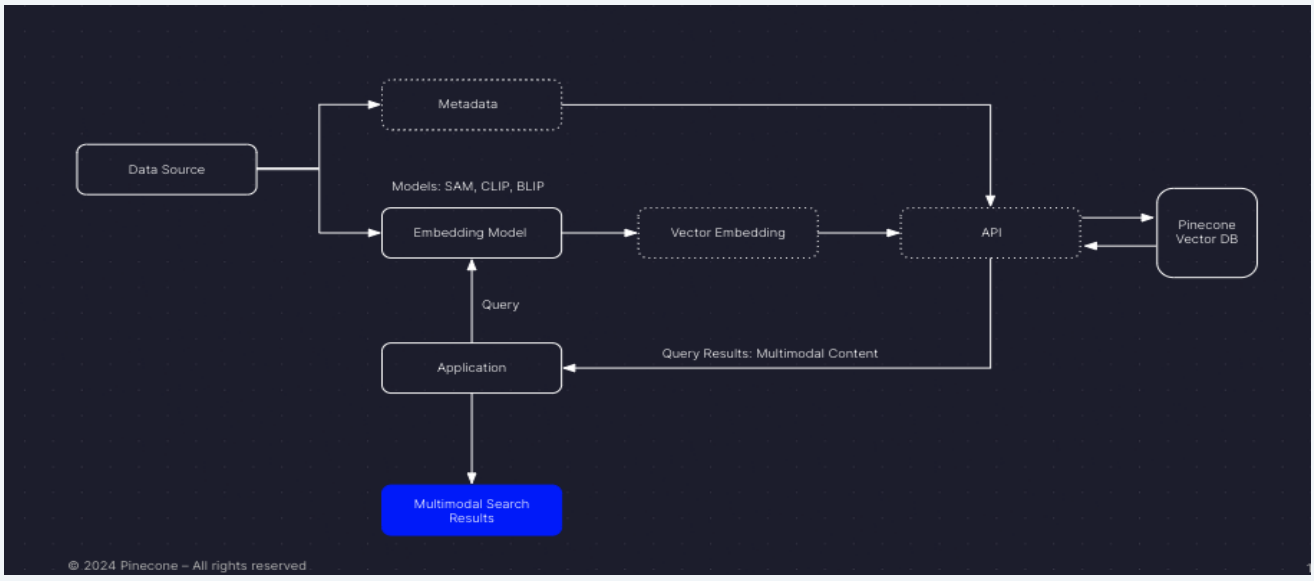


Figure 1. Multimodal search with Pinecone

Aligning images and text in a shared vector space

Transformer models and generative AI dramatically increase the ability to convert almost anything into embeddings. Approaches that use massive amounts of internet data have enabled a new frontier: aligning images and text in a shared vector space. Powerful models—such as Amazon Titan Multimodal Embedding, CLIP2 available in Amazon SageMaker, and BLIP—can understand and connect the underlying meaning or concepts across these various types of data. Images and captions, objects and attributes, and even audio and video can now be represented using the same high-level features.

Searching with a photo to find related products or news stories is one exciting potential of multimodal embeddings. Another is using a prompt to generate a dream home image that fuels a search through hundreds of thousands of listings to find a match.

This whitepaper explains the challenges of multimodal search with traditional databases and demonstrates why Pinecone is uniquely positioned to power multimodal use cases.

The possibilities of multimodal

Healthcare

Finding similar medical images and comparing them to a database of known conditions to diagnose a patient.

Education

Surfacing relevant learning and educational materials, such as images, videos, or audio clips, related to a specific topic or concept.

Media and entertainment

Using text queries, actor names, genre, or even an image of a favorite scene to search for movies or TV shows.

Advertising

Understanding the meaning and tone of the content and matching relevant advertising to it.

Travel and tourism

Planning trips on a travel platform that delivers destinations, attractions, and hotel packages using text queries, images, and even voice commands.

Multimodal with traditional databases can’t deliver accuracy or scale

Multimodal applications can be challenging to build using MySQL and NoSQL databases, which can’t search across different data types simultaneously. For text-to-image searches, developers manually tag each image with descriptive words. This is time-consuming and doesn’t yield accurate results.

Traditional image and video classification relies on labels and can find objects in media, but it fails to capture the semantic meaning of the media. Some developers use a traditional database for structured data and metadata while storing multimedia files on a file system or in object storage. Developers search for metadata using SQL queries and retrieve the corresponding multimedia files. However, the search is limited, it requires manual work, and it does not enable content-based retrieval.

Also, as datasets grow in size and complexity, maintaining real-time performance using traditional databases becomes increasingly difficult.

Vector databases improve retrieval and performance—but the right infrastructure is critical

A vector database indexes and stores vector embeddings for fast retrieval and similarity search, with capabilities like CRUD operations, metadata filtering, and horizontal scaling. These embeddings are a type of vector data representation that carries semantic information inside it. This information is critical in helping AI gain understanding and maintain the long-term memory needed to execute complex tasks. However, the scalability needed for modern multimodal search requires sophisticated and flexible infrastructure.

Pinecone yields accurate results at scale

Pinecone serverless is a managed, cloud-native vector database with a streamlined API and no infrastructure hassles. Pinecone serves fresh, relevant query results with low latency at the scale of billions of vectors and can handle large multimedia data volumes without compromising search speed or accuracy. It provides a unified and efficient solution and infrastructure for multimodal search, eliminating the need for complex hybrid approaches. You can access Amazon Titan models from Amazon Bedrock to convert data into embeddings and store them in Pinecone. And that makes it ideal for multimodal use cases.

Pinecone is purpose-built for multimodal search

Pinecone serverless is an excellent solution for any multimodal use case, but especially for those that require high performance, cost-effectiveness, ease of use, and high scalability. It supports vector embeddings from a wide range of state-of-the-art embedding models across multiple modalities, including text, image, and audio. This allows for powerful multimodal applications. Its advanced features, like filtering on metadata and sparse vector support, result in more precise querying and can boost relevance for multimodal searches.

Pinecone integrates with popular frameworks like Amazon SageMaker, LangChain, Haystack, Hugging Face, and more and data sources like Databricks, Snowflake, and Confluence. This allows it to fit smoothly into existing workflows and ecosystems for multimodal projects.

Pinecone in a nutshell

Pinecone's serverless architecture on Amazon Web Services (AWS) delivers many benefits and capabilities. These include cost reduction, scalability, multiple integrations, and improved performance. There are also helpful resources, an active community, and a plan that enables you to start for free.

High-dimensional data: Pinecone is specifically designed to store, index, and retrieve high-dimensional data as vector embeddings, which reduces latency down to milliseconds. Thanks to this low latency paired with high throughput for large datasets, Pinecone can support applications with up to billions of vectors.

Multitenancy: Namespaces, metadata filtering, and more enables you to partition your data, ensuring you're getting the right answers all the time at the lowest possible cost.

Scalability: Pinecone automatically scales based on usage, eliminating the need for manual capacity planning and management.

Vector clustering: Through proprietary algorithms, Pinecone uses vector clustering on top of AWS object storage. And with distributed storage, vector data is clustered on AWS object storage, providing virtually limitless data scalability and high availability.

Cost savings: Pinecone separates read and write paths, which allows the independent scaling of compute resources. Combined with its consumption-based pricing model, where you only pay for what you use, it delivers cost efficiency at any scale (up to 50x less than pod-based indexes).

Examples of multimodal search use cases with Pinecone

Among the number of multimodal search use cases, these are some of the most common:

Text-to-image search: Uses textual queries to retrieve images.

Image-to-image search: Finds images that are most similar to a given image.

Text-to-audio search: Searches for and retrieves specific audio content using text queries.

Text-to-video transcript: Searches and retrieves portions of transcripts.

You can also use combinations of these search methods. For example, you can augment an image-to-image search with text and make it (image+text)-to-image. The following chapters show you why Pinecone is ideal for these use cases and their combinations, starting with text-to-image search.

MULTIMODAL USE CASE

Text-to-image search

In a text-to-image search scenario, each image is converted into a vector representation using Amazon Titan Multimodal Embedding, CLIP2, or BLIP. A text query is converted into a vector representation using a text embedding model compatible with the image embedding model.

The Pinecone text-to-image advantage

Pinecone provides the vector database infrastructure to make large-scale, high-performance text-to-image search feasible and efficient. You can use it to search for the image vectors that are most similar to the text query vector, based on a similarity metric like cosine or Euclidean distance. The topmost similar images are retrieved (based on Top_k) and returned as the search results.

The Pinecone text-to-image search process

Let's say you're looking for a painting of the sun setting over the ocean from a collection of images. Here's the flow with Pinecone.

1. Image embedding

First, all source images are processed through a model like Amazon Titan Multimodal Embedding. The model generates vector embeddings that capture the semantic and visual content of each image. Those vector embeddings are stored in Pinecone.

2. Text query embedding

You can start with a text query like "a painting of the ocean at sunset." It is processed through the same multimodal model to generate a vector embedding that represents the semantic meaning of the text.

3. Vector similarity search

The text query embedding is then compared to all the image embeddings in Pinecone using similarity search. This finds the images with the embeddings that are closest to the query embedding in the vector space.

4. Results ranking

The most similar images are returned as search results, typically ranked by their vector similarity scores. Now you have the search results that closely match "a painting of the ocean at sunset."

That's how text-to-image search works. But what if you have a photo of the ocean at sunset and you want to find a painting that's similar? In the next section, you'll see how Pinecone can deliver image-to-image search.

PINECONE TEXT-TO-IMAGE SEARCH SCENARIO

Retail sales

A catalog and ecommerce retailer with a vast product image library has a Pinecone-based application that stores vector embeddings of the semantic and visual content of all the images. Team members find visuals using natural language descriptions. When preparing a campaign for eco-friendly outdoor gear, a search for “sustainable camping equipment in nature” instantly reveals a long-forgotten solar-powered tent and recycled material backpacks, reducing search time and surfacing underutilized assets.

MULTIMODAL USE CASE

Image-to-image search

In traditional image search engines, you typically use text queries to find images, and the search engine returns results based on keywords associated with them. In an image-to-image search, you start with an image as a query, and the system retrieves images that visually resemble the query image. You can also search for images that contain specific objects.

The Pinecone process for image-to-image search

When used for image-to-image search, Pinecone converts images into high-dimensional vector representations using an embedding model. It then indexes them for fast similarity search and retrieves the images most similar to a query image based on vector similarity. The process for this use case in Pinecone has a flow that's similar to text-to-image search.

1. Image embedding

First, all source images are processed through a model like Amazon Titan Multimodal Embedding. This model generates vector embeddings that capture the semantic and visual content of each image. Those vector embeddings are stored in Pinecone.

2. Query image embedding

When a user provides an image as a query, it is processed through the same model to generate its vector embedding.

3. Vector similarity search

Using similarity search, the query image's embedding is compared with all the image embeddings in the Pinecone database. This identifies images that have embeddings closest to the query image embedding in the vector space.

4. Results ranking

The most similar images are returned as search results, typically ranked by their vector similarity scores.

Image-to-image search with Pinecone can help companies across retail, fashion, agriculture, and other industries where image representation plays an important role, such as displaying products or identifying a disease from a photo of a symptomatic leaf.

In image-to-image search, a picture is worth a thousand words. However, there are times when a search for audio is just as valuable. In the next chapter, we show how Pinecone can also be used for that.

PINECONE IMAGE-TO-IMAGE SEARCH SCENARIO

Quality control

Multimodal search applications based on Pinecone can assist in quality control processes through the analysis of products or components images. By comparing images of defective and non-defective items, the application can identify patterns or anomalies that indicate potential defects. For example, when an inspector spots a potential issue, they can instantly query the database with an image and description. The system quickly retrieves similar past cases, helping identify defects more accurately.

MULTIMODAL USE CASE

Text-to-audio search

Text-to-audio multimodal search enables users to look for and retrieve relevant information from audio using written queries. A specialized audio embedding model converts audio files into vector embeddings in a high-dimensional space. A text embedding model converts text queries into vector embeddings and stores them in the same high-dimensional space. And this is where Pinecone comes in.

The Pinecone text-to-audio advantage

Pinecone's speed, scalability, and serverless architecture support text-to-audio search—especially on a large scale. Examples include looking for specific topics or quotes from a large collection of podcast episodes or querying recorded speeches or lectures from a conference for specific content.

The Pinecone process for text-to-audio search

The audio and text queries that are converted into vector embeddings are stored and indexed in Pinecone. Each vector is associated with metadata, such as the corresponding text segment, audio file, and timestamp. Pinecone's indexing process enables fast and efficient similarity search. When a user enters a text query, Pinecone performs a similarity search to find the text and audio embeddings that are most similar to the query vector. It uses indexed vectors and efficient search algorithms to identify and retrieve the top similar segments based on a cross-modal similarity metric, along with the corresponding text segments and timestamps.

Text-to-audio search with Pinecone can be used for analyzing television and podcast interviews, extracting summaries from important speeches, enabling lawyers to quickly find mentions of key evidence from recorded deposition transcripts, and more.

Now that you understand the role of Pinecone in multimodal search applications, let's look at how to build one.

PINECONE TEXT-TO-AUDIO SEARCH SCENARIO

Call center analysis

A customer support center has recordings of thousands of daily calls, but pulling common issues from them is difficult. Using a Pinecone-based system, each call is vectorized. When a new problem arises, such as a spike in “overheating” complaints, an analysis of related calls quickly identifies the source, and the support center takes action to solve it.

Building a multimodal search application using Pinecone and AWS

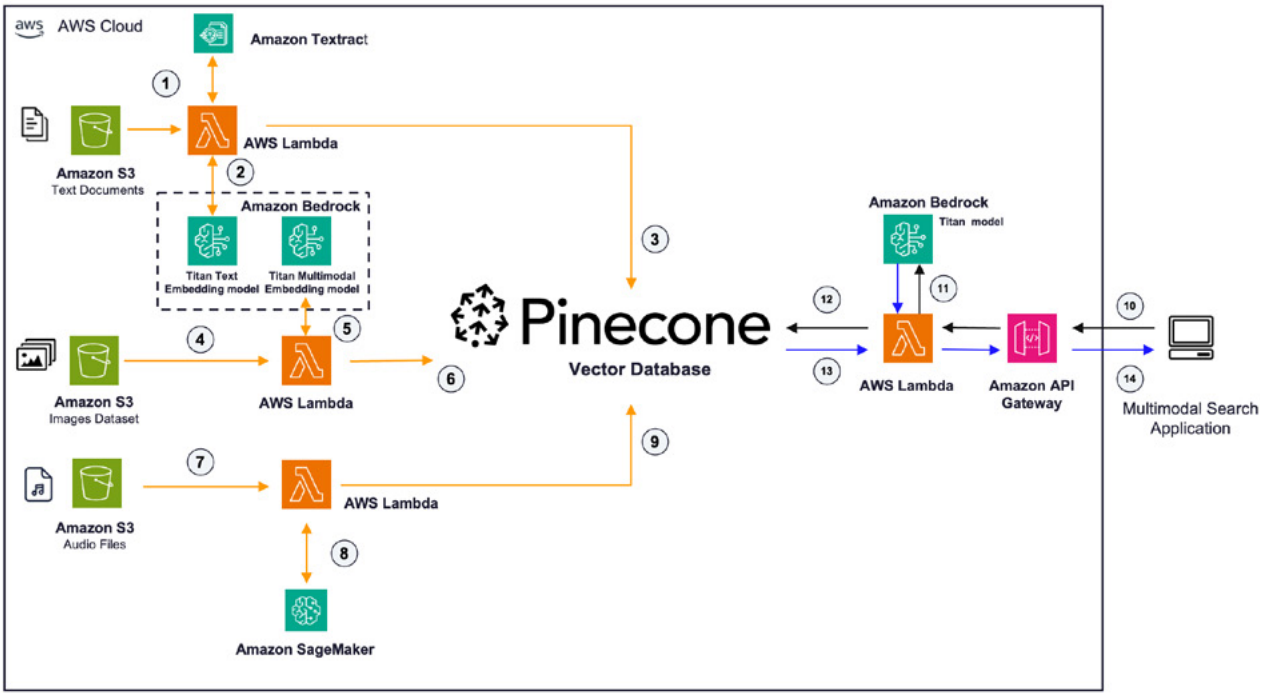
This chapter guides you through a reference architecture for building a multimodal search application using the Pinecone vector database, Amazon Bedrock, and AWS services. The architecture consists of two main components: the backend data pipeline and the frontend multimodal search application.

The backend pipeline converts multimodal data like text, images, and audio to vector embeddings and creates a centralized repository for them. It uses various embedding models from Amazon Bedrock to generate embeddings for text data and images, as well as Amazon SageMaker to host custom audio embedding models for audio files. These embeddings, along with their metadata, are stored in Pinecone.

The frontend multimodal search application accepts user search queries in multiple formats and converts them into vector embeddings. It then compares the query vector embedding with all the vector embeddings in the Pinecone database using similarity search. The most similar embeddings, along with their metadata, are returned as search results, typically ranked by their vector similarity scores.

In this reference architecture (Figure 2), Amazon Titan Text Embedding generates vector embeddings from text data, Amazon Titan Multimodal Embedding generates vector embeddings for images, and Amazon SageMaker hosts custom audio embedding models that generate vector embeddings for audio files.

Figure 2. Multimodal search application using Pinecone, Amazon Bedrock, and AWS services



Implementing the proposed solution involves two pipelines.

Backend data pipeline

The backend data pipeline is represented by numbers 1 through 9 in Figure 2:

1. Store the text documents in an Amazon Simple Storage Service (Amazon S3) bucket.
2. For each text document, use AWS Lambda to parse the text content from a simple file or a scanned document with Amazon Textract. Generate vector embeddings from the parsed text content using Titan Text Embedding in Amazon Bedrock.
3. Store the vector embeddings, along with their respective document metadata, in Pinecone.
4. Store the images dataset in an Amazon S3 bucket.
5. For each image, use AWS Lambda to generate vector embeddings with the Titan Multimodal Embedding in Amazon Bedrock.
6. Store the vector embeddings, along with their respective image metadata, in Pinecone.
7. Store the audio files in an Amazon S3 bucket.
8. For audio vector embeddings, Amazon SageMaker handles batch processing that allows for the generation of multiple audio files simultaneously. Use AWS Lambda to parse one audio file at a time and then generate vector embeddings with the custom embedding model.
9. Store the vector embeddings, along with their respective audio file metadata, in Pinecone.

Query pipeline

The query pipeline is shown as numbers 10 through 14 in Figure 2:

1. When a user provides a search query (it could be text, image, or audio), send it to the backend application via an API Gateway.
2. The search query is converted into vector embeddings using the respective embedding model. An AWS Lambda function orchestrates the process of converting the query to a vector, performing the search, and relaying the results.
3. Compare the query vector embeddings with all the vector embeddings in Pinecone using similarity search.
4. The most similar embeddings, along with their metadata, are returned as search results, typically ranked by their vector similarity scores.
5. Return the search results to the frontend application interface.

And that's how you can use Pinecone, Amazon Bedrock, and AWS services to build a multimodal application.

To learn more, visit _____ on GitHub.

Pinecone

Search for meaning everywhere with Pinecone and AWS

When you use Pinecone to add multimodal search capabilities to your applications, you get the freshest and most relevant results. Pinecone is fully managed, so it's easy to use, and there's no infrastructure to manage. When Pinecone and AWS are used together, you get more flexibility and capabilities—plus seamless integration with other AWS services—for optimal, accurate multimodal search.

Start building today with Pinecone or find Pinecone in AWS Marketplace.

aws Available in AWS Marketplace