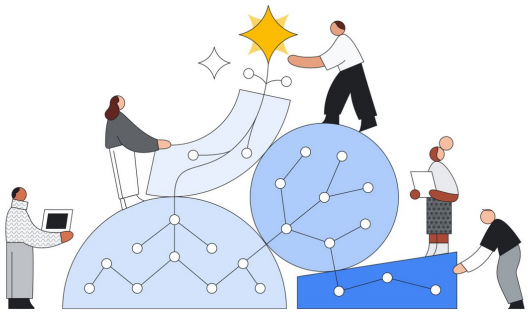
# Multimodal generative AI search | Google Cloud Blog

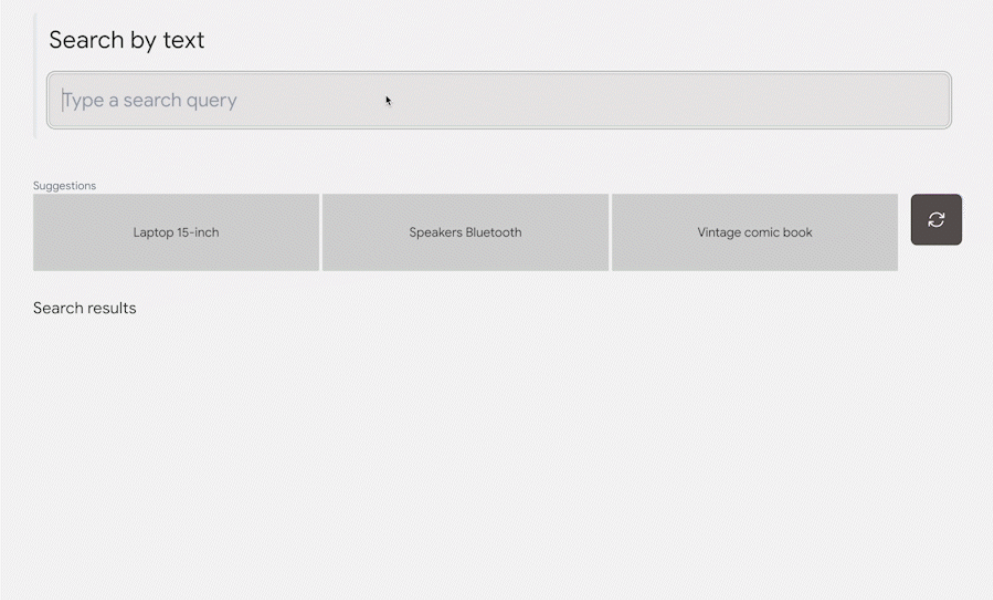




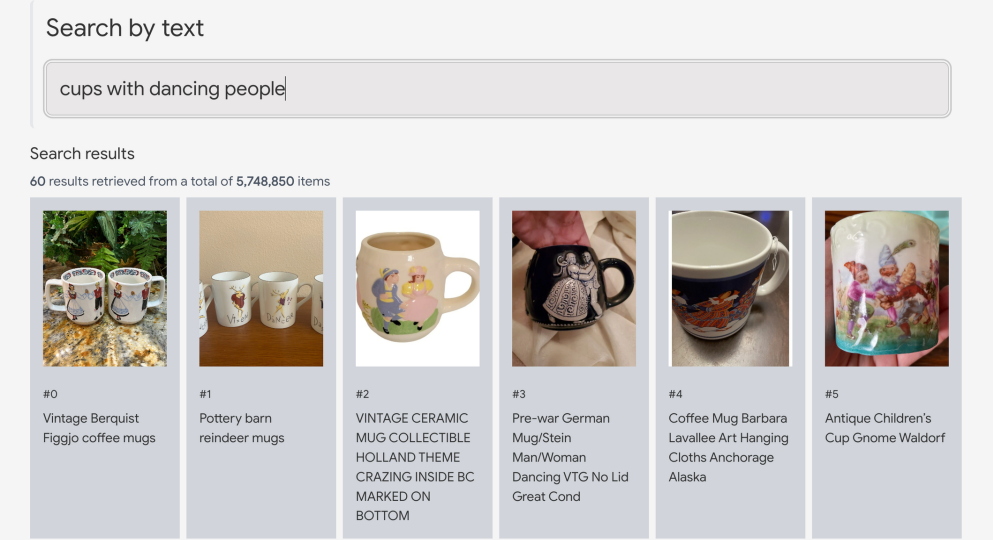
What if large language models (LLMs) had "vision", the ability to understand the meaning of images? Just like we have seen the innovation with LLMs with chatbots and text data, the ability would make another huge impact on businesses by letting LLMs look at and organize millions of images in enterprise IT systems. In this post, we will learn how a large vision language model (VLM) works and changes the business in the next couple of years.

Before we go into the details, you can experience the power of the VLM for yourself with this live demo. The demo was developed in collaboration with Mercari, a well-known marketplace app with over 50 million downloads in the United States. We imported 5.8 million item images from Mercari into the demo, passed them to Vertex AI Multimodal Embeddings to extract multimodal embeddings, and then built a search index with the embeddings on Vertex AI Vector Search. The sample code is also available here.

To interact with the demo, choose MERCARI TEXT-TO-IMAGE and enter any text query to find items. For example, a query for "handmade accessories with black and white beads" returns the following items, searched from 5.8 million Mercari product items in milliseconds. Please note that this demo does not use any of the item titles, descriptions, or tags for the search. The results are **retrieved by looking only at the item images** with the VLM.




A query result for "handmade accessories with black and white beads" on [Mercari text-to-image search demo]




A query result for "cups with dancing people"

The following are the points that make this demo unique:

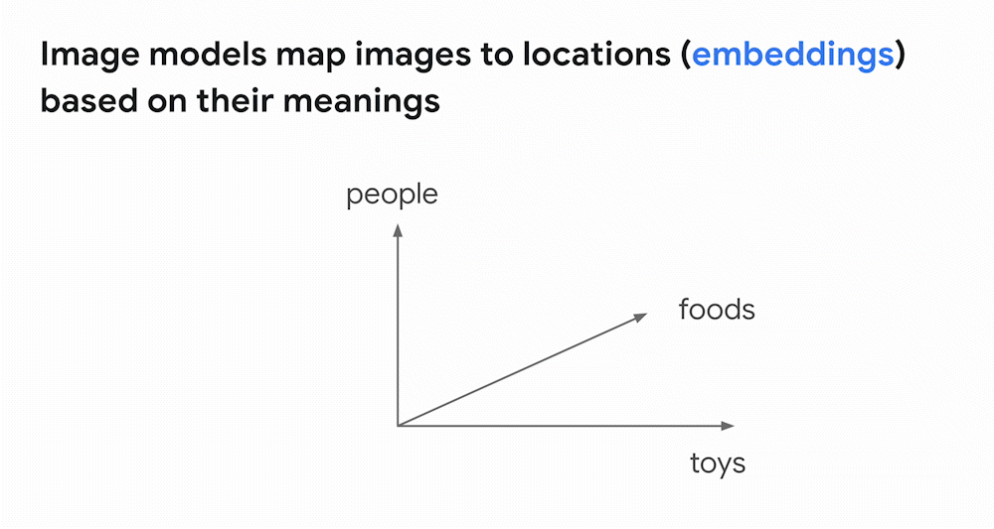
- **Multimodal semantic search with LLM intelligence**: Google Cloud launched [Vertex AI Multimodal Embeddings] early this month as

General Availability. The product uses the VLM called Contrastive Captioner (CoCa) developed by the Google Research team. In a nutshell, it is a vision model augmented with LLM intelligence that can look at either images or text and understand their meaning. In the "Cups with dancing people" example above, the CoCa model sees that the objects in the image are cups, and that there are dancing people drawn on their sides.

- **Grounded in business facts, scalable, fast and cost-effective**: This demo has the same design pattern and benefits as explained in the previous post Vertex AI Embeddings for Text: Grounding LLMs made easy. The search results are real business data available on Mercari, and no artificial texts, summarization, or images have been added. As a result, you can deploy this solution to production today without having to worry about the unexpected behavior of LLMs. The search results are returned in tens of milliseconds, without having to wait for sluggish text generation and incurring much higher costs.

## How does multimodal search work?

As discussed in the previous post, one of the most powerful applications of deep learning models is to build an embedding space, which is essentially **a map of meanings** for texts, images, audio, etc. For example, with an image model, images with similar appearance and meaning will be placed closely together in the embedding space. The model can map an image to an embedding, which is a location in the space. Therefore, if you look around the embedding, you can find other images with similar appearance and meaning. This is how image similarity search works.



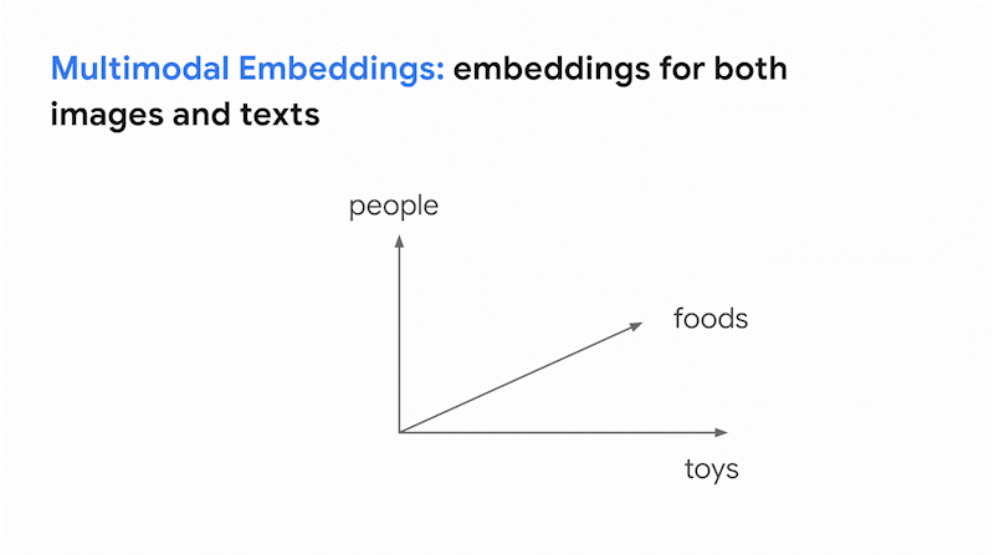

Similarly, a deep learning model can be designed to be trained on pairs of images and texts. The following animation illustrates how such a  model is trained. The model has three sub-models:

- A model to obtain image embeddings
- A text model to obtain text embeddings
- A model to learn the relationships between them

This is analogous to adding vision capability to a LLM.

The result is a vision language model (VLM) which can build **a shared embedding space for images and texts** with a fixed dimension, organized by their meanings. In this space, images and texts with

similar meanings are placed close together. This means that you can search for images based on text (text-to-image search) or search for text based on images (image-to-text search). This is the basic idea of how Google Search finds relevant results across images and texts.

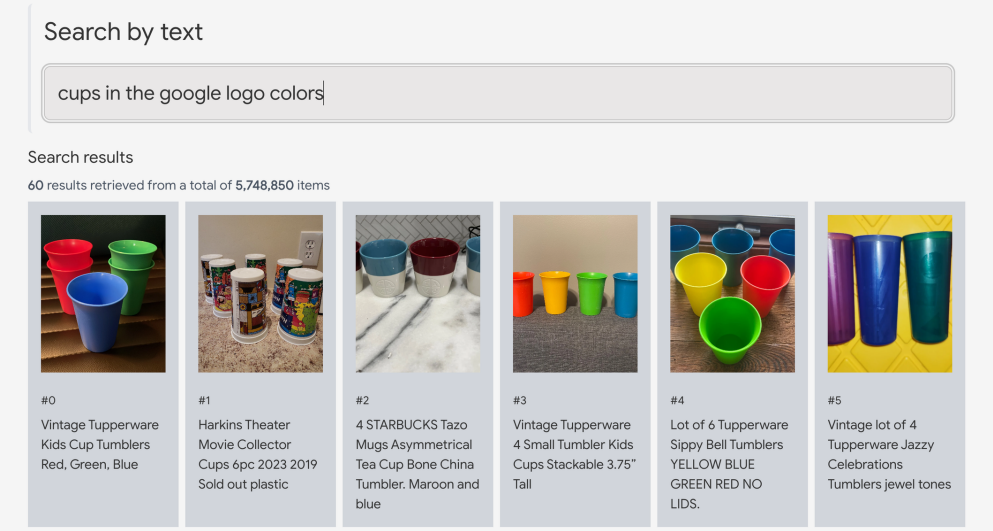



The innovative aspect of the recent large VLMs is not only their capacity to search across images and text, but also their **librarian-level LLM intelligence** in organizing them for various business use cases, without any effort for collecting industry-specific datasets nor additional training/tuning. This was almost impossible with the image-only deep learning models of the past.

How do the VLMs perceive the world of images? With the multimodal embedding space of 6 million Mercari item images, Nomic AI and Google have created a visualization demo that allows you to explore the space. The images are sorted into extremely specific categories, providing a glimpse into the complex way that model understands the images.

The multimodal embedding space with 6 million Mercari items (explore it here)

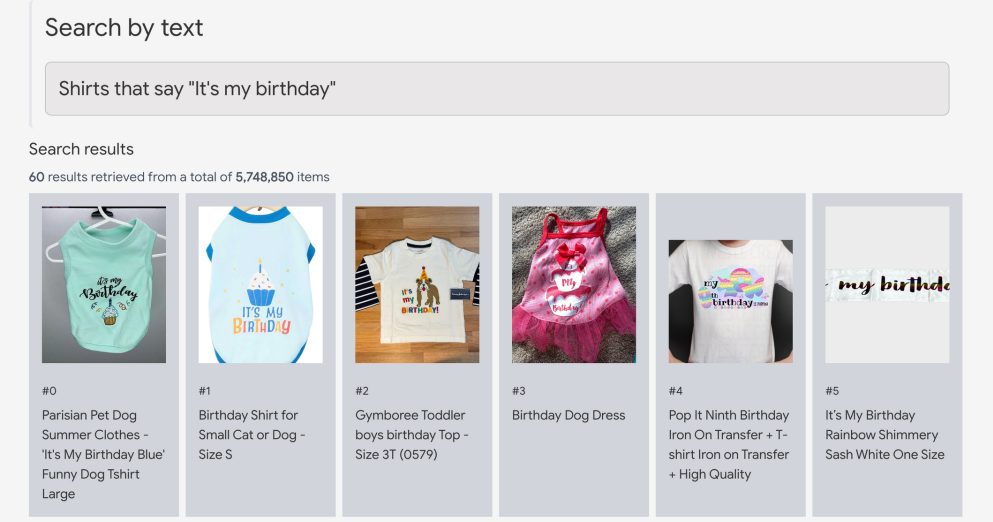
As an illustration of such intelligence, a query for "cups in the Google logo colors" on the multimodal embedding space returns the following results. The model can identify the colors of the Google logo and which images contain those colors, all without any explicit training (zero shot learning).




"Cups in the Google logo colors"

The model in the example below directly reads the text on the images without the need for any explicit optical character recognition (OCR)

process.




"Shirts that says my birthday"

## "LLMs with vision" change businesses

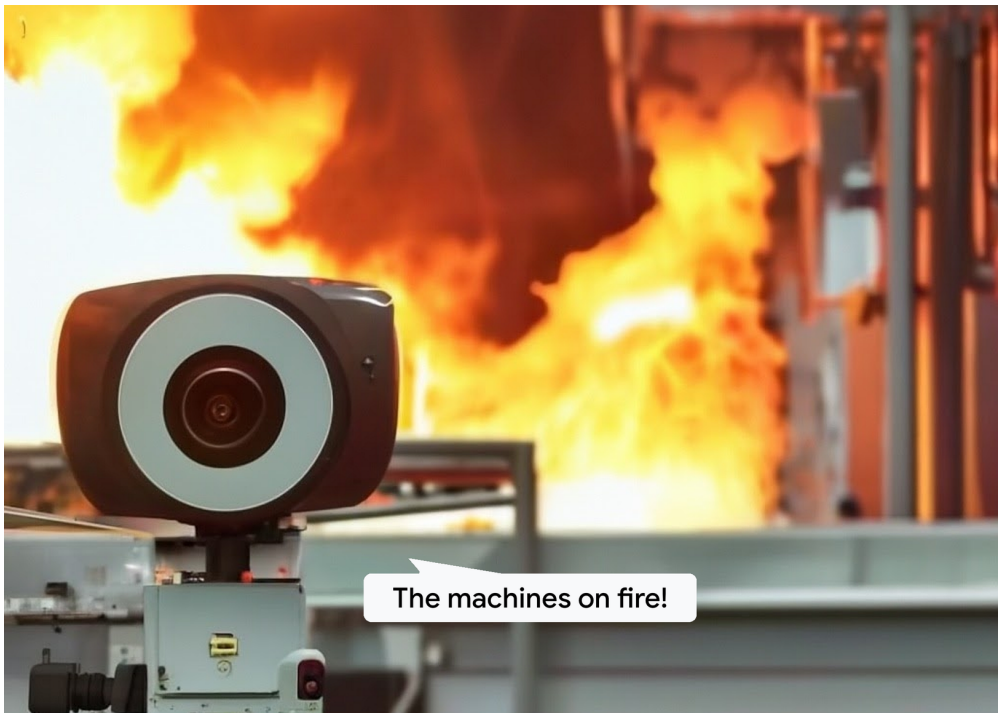
The exceptional performance of multimodal search enables a next-gen user experience in information retrieval across a variety of businesses, far exceeding the keyword search experience.

With **e-commerce** or **marketplace services**, sellers can upload an image of the item they want to sell. Then the service will search existing items with the same category or brand, or similar colors and styles. Using the Vertex AI PaLM API, the service can generate suggestions for the item name, description, and sell price based on the information it finds. This can significantly reduce the effort required for sellers to put articles on the market. Buyers can enter a natural language text query to find items that match their general intent, rather than just the exact item name.

LLM insights can make **security monitoring** much smarter for security camera images. Even with thousands of security cameras in the field and millions of monitoring images incoming every minute, you can still watch for images that match text queries like "a person trying to open the doors," "water is flooding in the factory," or "the machines are on fire."

**Autonomous driving manufacturers** store vast amounts of images and videos, which they use to train their vision recognition models. It

is critical for them to organize and retrieve these images quickly using complex query conditions, such as "a crossing road with red lights on and pedestrians are standing," or "a crushed car stopping in the middle of the freeway ahead" for the model development and verification. They can use natural language queries to find relevant scenes from millions of images instantly, even without any tags or labels attached on the images, significantly improving their productivity.




A possible use case: LLM-enabled security camera detects "the machines on fire" (AI generated image)

## Multimodal Search in Google Cloud

In Google Cloud, there are some ways to use multimodal search. Let's take a closer look at each option.

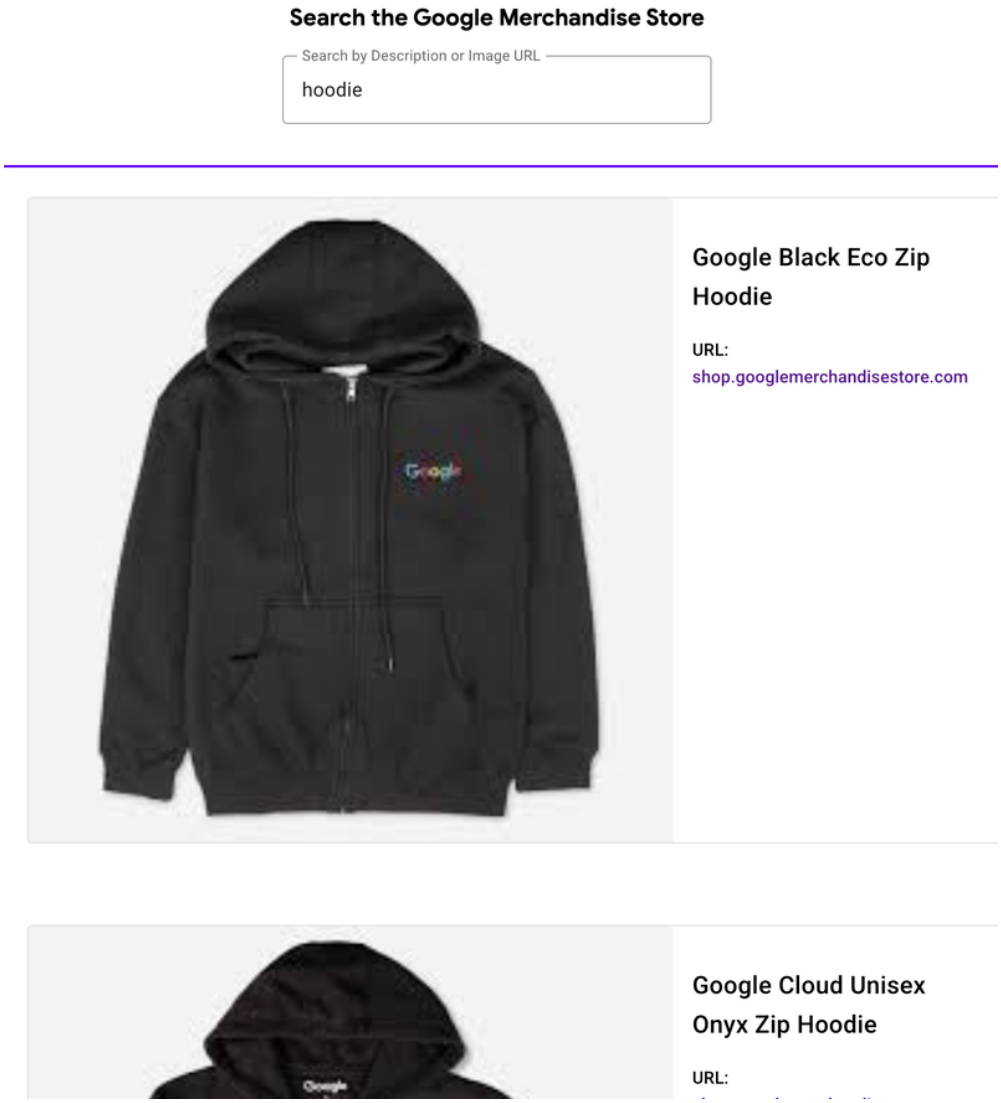
| | Option 1: Vertex AI Search (website/unstructured app) | Option 2: Vertex AI Vision Warehouse | Option 3: Vertex AI Search (structured app) | Option 4: Vertex AI Vector Search |
|---|---|---|---|---|
| Target developers | IT engineers, requires no ML expertise | IT engineers, requires no ML expertise | IT engineers, requires basic ML expertise | ML engineers/Data Scientists |
| Search algorithms | Keywords + embeddings-based search | Embeddings-based search | Keywords + embeddings-based search | Embeddings-based search |
| Vision Language Model (VLM) | Google Internal | Google Internal | Vertex AI Multimodal Embeddings (CoCa) | Vertex AI Multimodal Embeddings (CoCa) |
| Out-of-the-box solution | Yes | Yes | Yes (for the search engine) | No |
| Searchable data | Web pages/PDF files indexed on Search | Media (images and videos) indexed on Vision Warehouse | Tabular data indexed on Search | Any images or texts indexed on Vector Search |
| Embeddings are exportable? | No | No | Yes | Yes |
| Additional cost for creating embeddings? | No | Yes | Yes | Yes |

Options for multimodal search in Google Cloud

## Option 1: Search website/unstructured app

If your items are web pages or PDF files, Vertex AI Search is the easiest way to deploy and operate multimodal search capabilities for them. It is an out-of-the-box, fully managed search engine with lower integration and operation costs. All you need to do is create a website app (for web pages) or unstructured app (for PDF) on the console and specify the location of the contents (this video shows how it is easy to get started).

Once the index is ready, you can make a query with text or image to search for the pages and documents by their images. You can try it on this demo site. The search results will include not only image search results, but also a combination of keyword search and semantic search on texts and images, with a built-in ranking algorithm.




Example of image search results with Search website app (try the demo)

The caveats of this option are:

- Only web pages or PDF files are searchable
- The image embeddings cannot be accessed and reused for various purposes

The second point means that there is less flexibility for controlling the search functionality and quality compared to the other options.

## Option 2: Vertex AI Vision Warehouse

Vertex AI Vision Warehouse is an ideal product for those who seek to build a repository for multi-modal assets, such as images and videos and perform AI-based search on them, such as semantic search on millions of video clips for a video broadcaster, or similarity search on

product images for a retailer. Vertex AI Vision Warehouse users can build their own Warehouse, with an API-first approach and connect with their Vertex AI Vision applications.

Vertex AI Vision Warehouse is an out-of-the-box solution for image/video assets management with low integration and operation costs.

### Option 3: Search + Multimodal Embeddings

If your items are managed as tabular data in databases or storage, one solution for multimodal search on the items is to combine Vertex AI Multimodal Embeddings with Search structured app. In a case where you have product data in a database table and product images in storage, you can import the embeddings into the search engine with the following process:

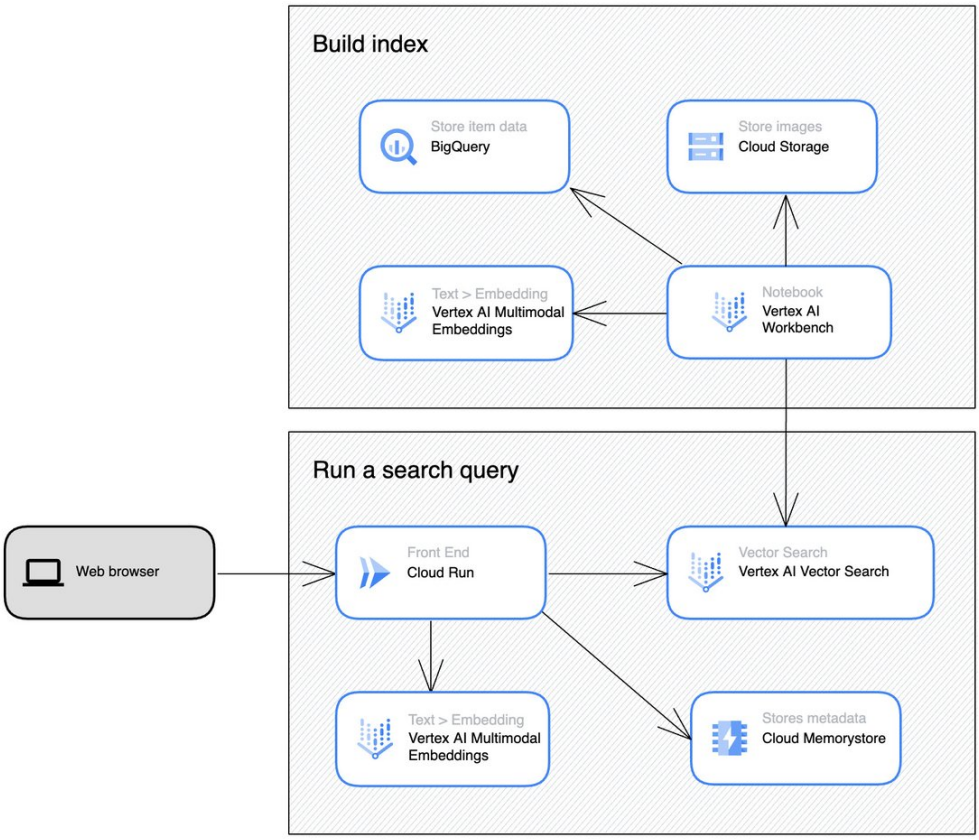
1. Pass the product images to the Vertex AI Multimodal Embeddings API to generate multimodal embeddings for each product
2. Import the embeddings as an additional column to the product table
3. Create a structured app of Search, specify an embedding config, and import the table
4. For each query, use the Multimodal Embeddings API to generate an embedding for the query text. Then, issue a query on the search engine with the query and its embedding

This feature is known as custom embeddings in Search structured app (currently in Preview phase). As with option 1, the search engine is out-of-the-box and fully managed, which reduces the cost of integration and operation. The search results will be a combination of keyword search on product attributes and multimodal semantic search on product images.

### Option 4: Multimodal Embeddings + Vertex AI Vector Search

If you have a team of machine learning engineers and would like to have full flexibility in designing multimodal search and reusing the embedding for various purposes such as recommendations, then a

combination of Vertex AI Multimodal Embeddings and Vertex AI Vector Search is the correct choice. The Mercari image search demo shown earlier was built using this approach.




Mercari multimodal search demo architecture

Vector Search is a bare-bones vector search engine. It can be used to build your own search or recommendation product from scratch, requiring ML expertise. The search latency is as low as tens of milliseconds in many cases, and also you can specify how the index and its shards are configured to optimize the search quality and latency. Vector Search also supports stream update to add or update each embedding in the index in real-time. This is suitable for cases where you need to add or update the items frequently and reflect it to query results in a few seconds.

Search is a full-fledged search solution that encapsulates a wide variety of functionalities, such as token-based search with keywords, vector search with embeddings, with tokenization, spell correction and synonyms, and sophisticated re-ranking algorithms and personalizations. It is provided as an out-of-the-box, fully-managed service, and is best suited for IT engineers who want to minimize the

cost of integration and operation with less ML expertise. As it handles the complex query and index processing under the hood, the latency for query and index update is relatively higher than Vector Search.

Both products are capable of performing LLM-enabled multimodal search. However, it is important to understand the characteristics of each product to make the best choice for your needs.

## Get Started with multimodal search

The innovation of LLMs is not limited to text chat use cases. Multimodal search is just as powerful, with the potential to significantly improve business productivity. Let us begin using this technology with Google Cloud tools.

### Resources

#### Demo and sample