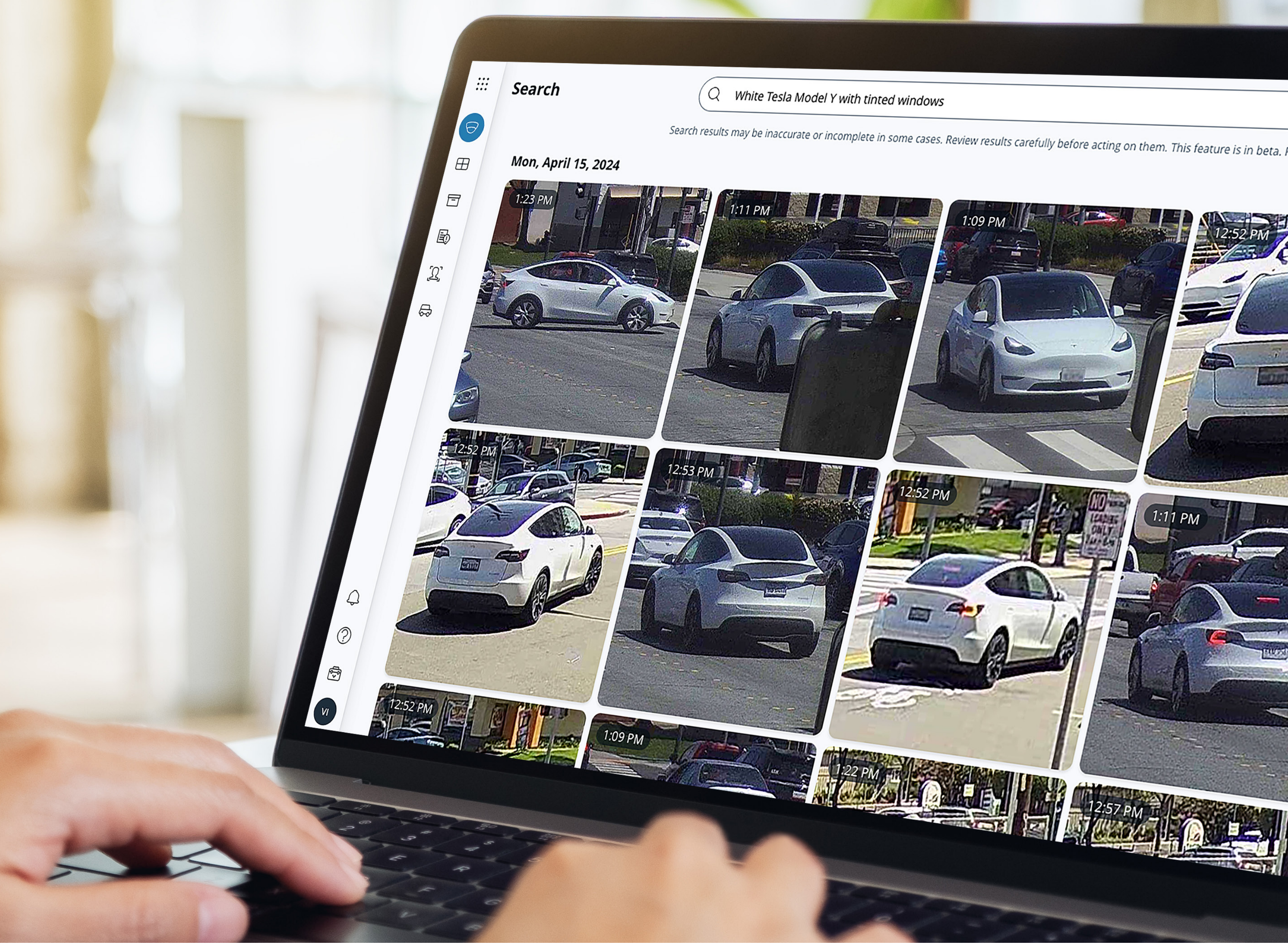


Search results may be inaccurate or incomplete in some cases. Review results carefully before acting on them. This feature is in beta.

Verkada

AI-Powered Search at Verkada

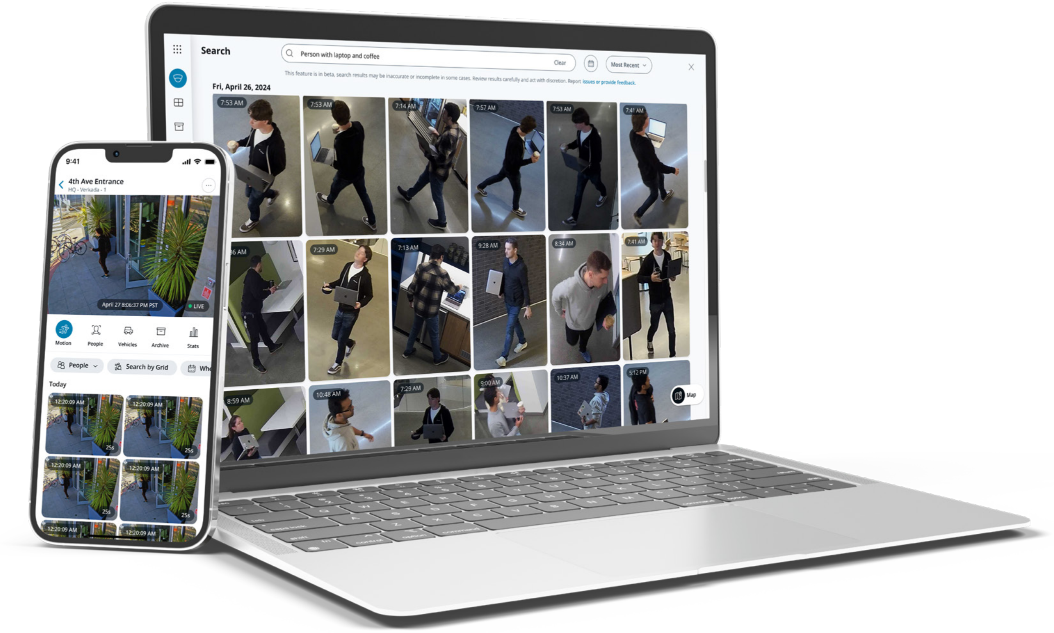
Search at Verkada

Building a Hybrid Cloud Foundation for AI-Powered Investigations

Introduction

At Verkada, we often speak of our hybrid cloud approach to physical security and its benefits for storage, bandwidth utilization, processing, security, and data privacy. This foundational architectural decision–one where we store and process data both on our cameras and in backend data centers–has unlocked different computer vision (CV) analytics for our customers. Our camera search functionality today–AI-powered search for granular attributes of people and vehicles–has come a long way from Verkada's inception. Our original decision to offer customers a hybrid cloud system has, though, enabled us to continuously innovate from the start and, eventually, arrive at our notable AI-powered capabilities today.

In this white paper, we'll begin with an exploration of our hybrid cloud architecture and illustrate how processing both on cameras at the edge and at AWS' backend data centers evolved our CV analytics from basic person detection to today's AI-powered search–highlighting the importance of dual processing, flexibility and scale in supporting these CV features. We will then dive deeper into our choice of foundation AI model, and why that's not only the most suitable model for our customer's needs today, but how we've improved on this foundation model in important ways. It will be clear that, just as our CV capabilities began with limitations, our AI-powered search has its own set of limitations. Building this AI functionality on our hybrid cloud architecture, however, will similarly allow us to continually improve on its functionality. We'll then conclude with an overview of our moderation techniques–or, how we're tackling AI-powered search with social responsibility, respect and anti-bias.

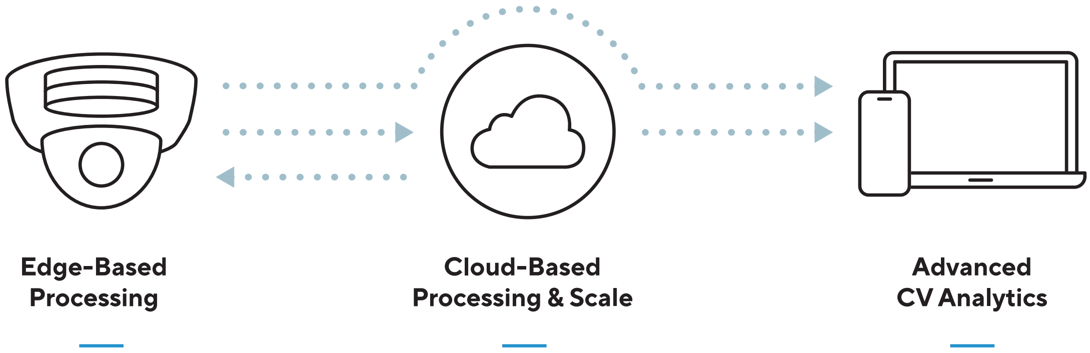


This feature is in beta, search results may be inaccurate or incomplete in some cases. Review results carefully and act with discretion. Report issues or provide feedback.

[1] Check out our User Guide for example queries and methods for constructing your own queries relevant to your organization's requirements.

Hybrid-cloud architecture

A Future-Proof Choice for AI



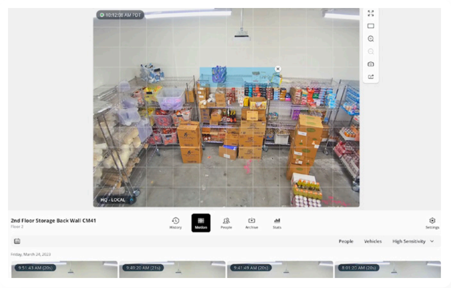
When we set out in 2016 to explore a new approach to physical security management, several architectural choices existed: a traditional, closed-circuit NVR/DVR-based system held the promises of airgapped security, but fell short on scale and distributed management. A pure cloud system offered the necessary scale and distributed, remote management, but constantly streaming video to the cloud presented serious bandwidth issues–compromising video quality and retention. We opted instead for the best of both: a hybrid cloud system where we could store and process video both on the cameras themselves and over the cloud–specifically using AWS because of its notable data protection and privacy, security, and compute at scale. The benefits of a hybrid cloud security system are now well-understood and a handful of security systems vendors utilize this approach today. Our early decision to ground our systems in a hybrid cloud architecture, though, has enabled us to innovate quickly and incrementally improve our analytics at scale–results that would have proved difficult or impossible without a hybrid cloud system.

Verkada's cameras include on-device Ambarella processors, with additional processing and compute over AWS. This initial hybrid decision grounded us in a mindset of how to offer the promises of a hybrid cloud system to customers for their investigations and near real-time monitoring. Our first step, then, was to leverage hybrid cloud processing to help customers search for motion events–typically the most relevant types of events in security incidents.

The timeline below depicts Verkada's initial years where motion search (i.e., basic object, people, and vehicle) and attribute detection relied mainly on backend processing–limiting the power of video search and attribute detection.

Our early days

Exploring the Limits of Edge-Based Processing and Backend Compute

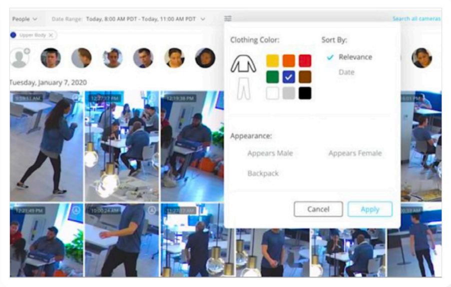


HiSi processors on Dome Cameras

2016-2017

Verkada's founding and the first iterations of motion detection on Command

- Similarly to today, our cameras uploaded thumbnail images once every 20 seconds to AWS to optimize bandwidth consumption. Unlike today, however, where we support near real-time motion detection and trajectory tracing, our camera thumbnails were divided into a 10x10 grid (consisting of 100 sub-blocks) and, if we detected pixel changes moving from one sub-block to the next, we would capture a thumbnail of that motion change (in a "motion thumbnail") and send it to AWS for additional processing. At the time, though, "motion thumbnails" could only be sent every 10 seconds and, once triggered, could not be sent again until the next 20 second interval. Of course, motion in a camera's field of view can occur again in under 20 seconds (i.e., a second object enters that camera's field of view during the period that the first "motion thumbnail" is undergoing processing), resulting in our cameras' inability to detect a second or third moving object–presenting obvious limitations for our customers.

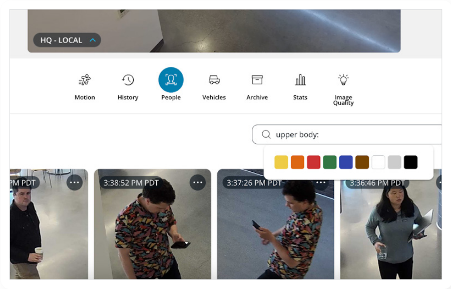


Ambarella CV2x processors on Dome, Mini and Bullet Cameras

2018-2019

Integrating third-party APIs for CV analytics

- On-edge object detection (Single Shot Detector API), on-cloud attributes detection (Sighthound API), face search (AWS API)
  - In addition to motion search, our hybrid cloud also supported more refined CV capabilities like people and vehicle attribute detection. Here, though, we utilized a series of third-party APIs and AWS models which primarily relied on backend processing, increasing latency and drawing substantial compute from AWS.
- It was around this point in time where we developed the concept of a "hyperzoom," (HZ) or a highly-detailed segment of a high-resolution image that contains only a person or a vehicle–items we felt most imperative to an investigation. Instead of running the attribute model on an entire thumbnail of footage, we ran it instead on a HZ–reducing search latency and improving investigation accuracy and relevance. Our initial attribute model allowed users to search on HZs of people and vehicles using a predefined list of search characteristics. Complementing the CV2x's accelerated processing of HZs at the edge, the HZ was then sent to AWS' backend data centers, which hosted our attribute model for scale, security, and data privacy–an example of how our hybrid cloud architecture began supporting enhanced CV analytics.



2019

Improving CV analytics with third-party APIs

- Cross-camera person / vehicle attributes search (Sighthound API)
- Profile matching (AWS API)
- Person of Interest (AWS API)

1. With edge-based processing still in its early stages, we leveraged third-party APIs that mainly relied on backend computing like Single Shot Detector for object detection and Sighthound for attribute search.

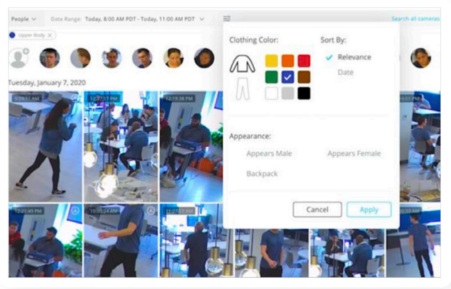
Realizing the benefits of hybrid cloud

Supporting a Natural Technological Progression

Conducting effective investigations, however, required near real-time, multi-object motion detection and detailed attribute search. With time, processors and backend hosting improved and the promises of hybrid compute became clearer. In tandem with these developments, we continued to invest in CV innovation internally–beginning instead to build attribute search, object detection and motion models in-house and migrate some processing from the backend to the cameras themselves.

As the timeline highlights below, we've leveraged our hybrid cloud foundation to complement agile edge processing with scalable in-house models hosted on AWS to offer our customers new, improving features:

Establishing the modern elements for AI-powered search

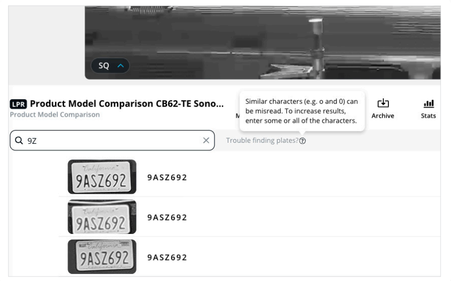


New Ambarella CV2x processors

2021

Begin migrating CV models in-house for advanced analytics like attribute filters on people and vehicles, license plate recognition and occupancy trends

- In-house person and vehicle attribute model



Product Model Comparison CB62-TE Sono...
Product Model Comparison

Similar characters (e.g. o and 0) can be misread. To increase results, enter some or all of the characters.

9Z

9ASZ692

9ASZ692

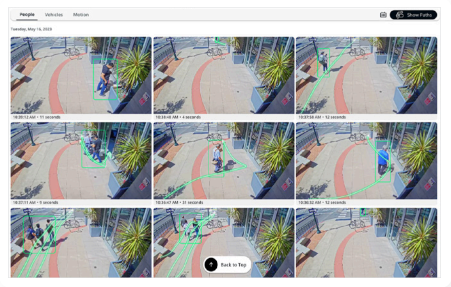
9ASZ692

Ambarella CV2x processors

2022

Building on our new in-house model paradigm

Creating our POI model, face search model and tracking-based license plate recognition



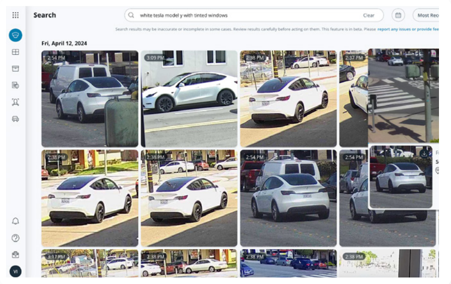
Extending the Ambarella CV2x processor to Dome, Mini, Bullet, Fisheye, Multi-Sensor, PTZ

2023

Continuing to improve on analytics and motion search in-house

- Edge-based processing using the CV2x allowed us to break down video and analyze people and vehicles at a rate of 10 frames per second. From this person and vehicle detection, we could create "tracklets," or associations of these objects across space and time. From these tracklets we could extract the most informative HZ (i.e., high resolution, visible facial features, specific vehicle criteria) for more advanced analytical purposes.
  - **MotionX:** With the "tracklets," we could then recreate the entire trajectory of a person or vehicle within a camera's field of view and trace motion of an event. MotionX trajectory analysis constitutes one of several CV features our hybrid cloud makes possible.

- **Polygon-based motion alerts (line crossing, crowding and loitering):** line crossing, crowding and loitering alerts: Leveraging our hybrid cloud, we could also incrementally improve our motion detection to dramatically reduce the multi-object blindspots of our early years. With the CV2x, we've evolved from creating a "motion thumbnail" once every 10 seconds, to identifying motion on 10 frames every second–opening up near real-time compute of people and vehicles at the edge. With this advanced on-camera processing in place, we could also buttress our 10-frames-per-second edge processing with a variety of in-house models: building capabilities like person or vehicle trajectories, near real-time person of interest (POI) alerting, line-crossing, and crowding/loitering alerts on our platform.



Search results may be inaccurate or incomplete in some cases. Review results carefully before acting on them. This feature is in beta. Please report any issues or provide fee

Fri, April 12, 2024

Extending the latest Ambarella processors to all cameras

2024

AI-powered search

The culmination of years of innovating on a hybrid cloud architecture

The next logical step: how our hybrid cloud architecture supports AI-powered search in Command

We've now arrived at the next stage of this technological progression: incorporating the latest in AI into our search capabilities. The confluence of our Ambarella processors delivering near real-time edge processing, our bandwidth-friendly upload capabilities and the robust compute, scale and security of AWS on the backend collectively give us the ability to not only support massive open source AI models, but also modify (and improve on) them in-house. This approach should come as no surprise as it follows our historic product development cadence that we outlined above: improving on open source tools in-house by tailoring them to our customers' requirements.

Our foundation AI model of choice: one with top "zero-shot" performance on millions of HZs

Our current form of attribute search leverages hyperzooms (HZs) of people and vehicles and limits our users to a predefined list of search attributes, like clothing color or vehicle type. Now, with the advent of AI-based image analysis, we can run these HZs through an open source model for processing and pair them with freeform queries–expanding our users' search capabilities from a predefined list to near-limitless search criteria. Selecting the right foundation model to support these ambitions, however, required careful consideration and for our customers, we chose an open source implementation of OpenAI's Contrastive Language–Image Pretraining (CLIP) model.

CLIP allows us, in short, to pair one's natural language with the most likely corresponding set of images. CLIP, created in 2021,[2] currently represents one of the most efficient and accurate models of its kind when it comes to "zero-shot" performance–or, how it matches text to images not in its original training set.[3] It's crucial we offer an AI search functionality with top "zero-shot" performance as, of course, none of our customers' HZs were part of CLIP's training.[4]

[2] CLIP's origins are also derived from thorough academic research on text-image machine learning models.
[3] See the "Key Takeaways" section here for statistics on CLIP's efficacy and "zero-shot" accuracy relative to other peer models.
[4] See here for a detailed explanation of CLIP's pre-training and "zero-shot" performance.

Why are we using CLIP for our AI-powered search?

CLIP constitutes a powerful and efficient model and outperforms many other peer models, especially as it pertains to zero-shot performance. When our customers use our AI-powered search, we'll process their HZs through CLIP–images that were not part of CLIP's training. As such, it's imperative that we deliver our customers a feature grounded in top zero-shot performance.

The massive training dataset and multi-modal training methods (image and text encoding) used to develop CLIP enable it to outperform many other models, both in terms of accuracy and efficiency. In fact, the open source version of CLIP that we're using for our purposes as a foundation model–OpenCLIP–is trained on over 2 billion text-image combinations and operates on a 2.5 billion parameter neural network, making it even more powerful for zero-shot results than the initial iteration of CLIP. Importantly for our customers, CLIP also outperforms many other models on a variety of granular zero-shot performance factors like text recognition (e.g., "FedEx truck"), image context (e.g., "person walking down a stairwell") and its ability to handle more complex query inputs than other models given its extensive pretraining on textual descriptions.[5]

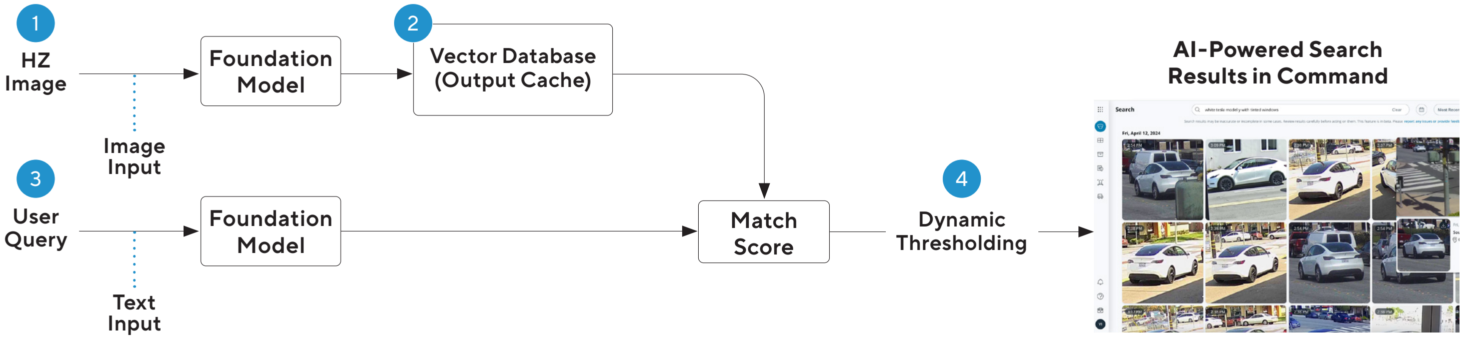
A continuation of our past: how we're improving on our foundation model in-house to directly address our customers' needs

It's evident why we chose OpenCLIP as our foundation model, but we had to adapt this foundation model's capabilities to the direct needs of our customers. In short, our foundation model is an image-to-text matching tool–not a search engine. We've, in turn, created a search engine out of natural language queries, with the foundation model as our "translator," probabilistically matching user query to image. This approach should come as no surprise–as with all of our CV capabilities, we utilize our hybrid cloud to build and support added functionality that directly benefits our customers. The foundation model is strong as an out-of-the-box matching tool, but to handle the millions of HZs from our customers on a daily basis, we've built on this foundation model in ways pertinent to the security and operational needs of our customers.

Each day, our systems ingest vast amounts of footage with millions of captured images. If we were to process these images through the foundation model, the latency would prove too great for a seamless customer experience. Some queries would, too, display irrelevant or inappropriate results. We've thus employed a four-pronged approach–that all work in consonance on our system's backend and also on the user's frontend queries–to tailor the foundation model to our customers' numerous and varying requirements, as highlighted in the diagram below.

[5] See the "Key Takeaways" section here, including specifically the statement that "[t]he best CLIP model outperforms the best publicly available ImageNet model, the Noisy Student EfficientNet-L2, on 20 out of 26 different transfer datasets we tested."

Our four in-house enhancements to the foundation model, which turns the foundation model into a low latency, scalable and pertinent experience for our customers conducting investigations.



1.

We only run HZs through the foundation mode

In line with our history of attribute searches, when a user inputs a query into our AI-powered search, we only run the HZs of people or vehicles through the foundation model, not the entire frame. With the large amounts of footage to ingest and then send to AWS for additional backend processing at scale, it's imperative that we minimize latency and maximize accuracy and relevance for our customers.

2.

Cache of model outputs

With the foundation model's several billion parameters, running each HZ through the foundation model to return the most relevant results requires massive compute and scale. While AWS supports the compute and scale requirements for this image retrieval process, it has the potential to take too long–compromising our customers' ability to run investigations quickly. We accordingly cache model outputs for the most recent 10,000 HZs on AWS so we don't need to run each HZ through the foundation model each time a user queries. When a user then queries in AI-powered search, we utilize our cache–rather than running all HZs through the full foundation model–to return the most relevant results. This process enables far quicker retrieval of results, at scale. See our FAQ here to understand how we're protecting customer data in this caching process. It's critical to note that we **do not** use customer HZs to improve our feature's functionality.

3.

Query matching: processing and parsing

When a user queries, we do not run that query verbatim through our foundation model or supporting cache. We will, for instance, add more specificity to a query so our models return more relevant results. We will append certain clarifiers like "a photo of," "a photo with," "an example of," or "a photo containing" to a query for greater matching accuracy as the foundation model was trained on captions with these phrases (e.g., a query for "person holding ski gear" might, on the backend, be amended to read "a photo with person holding ski gear"). Importantly, too, we've included parsing functionality that allows our users to search by time and space, so users can query for incidents that occurred at specific points in time, on specific cameras or sites.

4.

Dynamic thresholding

Of course, when we compare a user's natural language query to the result from the foundation model, we strive for accuracy–we want our users to receive the likeliest set of matches to their queries. Creating the foundation model involved grouping like images together into descriptor buckets (e.g., putting all images of dogs into a "photos of dogs'' category) and "penalizing" unlike images. This process worked on a scoring system where images that more closely matched the text descriptors received higher scores and those that didn't match received lower scores. At Verkada, we'll assess the distribution of scores for a given query's results, dynamically set a range of scores that correspond to the images' results that are relevant (i.e., an accuracy range) and only show our users results that fall into that specific, high confidence range–improving accuracy and relevancy of retrieved results.

Limitations and current results

While we currently leverage OpenCLIP as our foundation model, and we've constructed important algorithms on top of it (like HZ-only processing, cached retrieval, processing and parsing, and dynamic thresholding), this is our first iteration of AI-powered search. Given our hybrid cloud foundation and flexible backend processing, we're able to change our underlying model and algorithms should a different open source text-image pairing model present itself as more suitable for our customers.

In this first iteration of AI-powered search, customers may experience incorrect or incomplete results, which we are working to improve, both for recall and precision.[6] We've opted for a model that has higher recall than precision.[7] As a result, users will see all likely results of their investigation (inclusive of true positives and some false positives), and can determine the accuracy of their search results themselves. If we opted for a model with higher precision, some results identified by the system as false negatives would not be displayed to the user, which could be more costly than a false positive for an investigation. As designed, we put the customer in control to filter out false positive results for themselves.

Moderation

Unlike our filter-based attribute search, which has a predefined set of descriptors for narrowing search results, our AI-powered search lets customers search their footage for a wide range of attributes using their own words. Because our foundation model has been trained on publicly available data from the Internet, results may still be inaccurate, inappropriate or offensive. We have built query moderation into our platform to help reduce the risk that our AI-powered search is used maliciously or in ways that may be harmful. Moderation is a critical safeguard for this powerful feature.

At the same time, we implemented moderation in a way that isn't overly restrictive to the point that it compromises the feature's usability. We also leverage industry recognized practices, including open source data and OpenAI's moderation APIs, to make moderation more effective. Striking the right balance between respectful and useful searches is paramount for us.

Our approach to moderation

Just as we've chosen to use a publicly available model for our AI-powered search feature, we developed our query moderation using publicly available practices for effective moderation. OpenAI, for instance, has published guides and blogs for developing suitable moderation techniques.[8] We further augmented these capabilities with additional proprietary protections, which include our own list of prohibited search categories:

[6] In machine learning modeling, "precision" is the percentage of true positive results out of the entire batch of "positive" (inclusive of true and false positives) returned results. It's expressed formulaically as [Precision = True Positive Results / (True Positive Results + False Positive Results)]. It's a measure of the model's accuracy in identifying positive results. "Recall," on the other hand, measures the percentage of true positives that are correctly identified. It is expressed formulaically as [Recall = True Positive Results / (True Positive Results + False Negative Results)]. Whereas precision measures the overall rate of correct predictions, recall is helpful to identify the rate of false negatives.

[7] Note that with Large Language Models (LLMs), recall and precision statistics will differ depending on the measured sample sizes. Our model's average recall is greater than its average precision.

[8] See blog: https://openai.com/blog/using-gpt-4-for-content-moderation. For more detailed, technical approaches to moderation that we leveraged, see: https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf and https://cdn.openai.com/papers/GPTV_System_Card.pdf.

- Information about the race, ethnicity or nationality of a person
- Information about the educational qualifications or religious beliefs of a person
- Subjective descriptions of people (e.g., attractive, ugly, wealthy)
- Inferred content from an image (e.g., "coolest person in the world")
- Sexual content or innuendo
- Names of specific people (i.e., public figures)
- Inappropriate or offensive descriptions of people or (e.g., "dumb person")
- Slurs equating people with animals (e.g., monkeys)

We also tapped into open source data to generate a list of banned text strings, in languages with latin characters like English, Spanish, French, German, etc.[9] When words or phrases in a query appear on the list, we reject the search and mask any results–a process known as string-based matching. We also complemented this technique with OpenAI's moderation API to cross-check queries that create harmful or biased results.

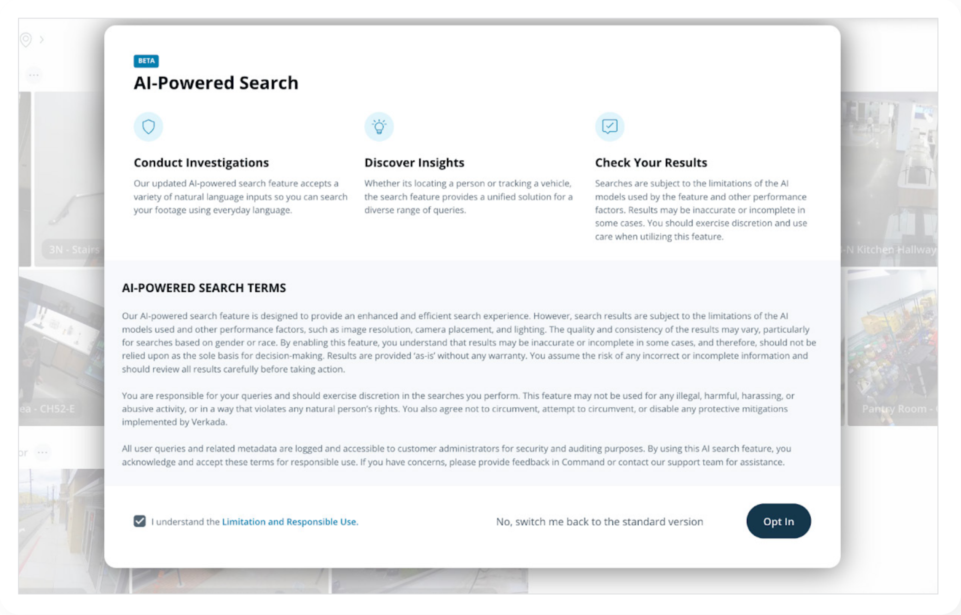
We also hold frequent engineering "bug bashes," analyze thousands of our own engineering test queries (i.e., classify thousands of queries as "appropriate" or "inappropriate") to modify our moderation rules and optimize them for usability and to avoid bias. We can update our "blocklist" in minutes and thereby continuously improve our moderation techniques.

Combining multiple open source models and in-house finetuning has, though, led to a moderation model that outperforms both Mistral-7b and ChatGPT 4 (two open source AI models with moderation measures) in our testing, both in our testing, both in terms of precision and recall:

	Precision	Recall
Mistral-7b		
Inappropriate	**31%**	**87%**
Appropriate	**98%**	**77%**
GPT-4		
Inappropriate	**58%**	**67%**
Appropriate	**96%**	**94%**
Our results: in-house finetuning of Mistral-7B		
Inappropriate	**96%**	**79%**
Appropriate	**98%**	**99%**

[9] Note: currently our AI-powered search only supports queries in languages with Latin characters.

While our moderation performs well, it is not perfect, and it is entirely possible that our users might see results that are inaccurate, incomplete, or inappropriate. We're committed to transparency and making these limitations clear to our users by including robust in-product disclosures that the users must acknowledge and agree to before they can begin using our AI-powered search (depicted below). As we continue to improve upon our moderation processes, we will continue to publish our results and additional methods we employ to reduce inaccurate, incomplete, or inappropriate results. In fact, we have built-in features for users to offer feedback on the quality of their results and organization administrators can review queries in their audit logs to ensure that no users have attempted inappropriate queries.



AI-Powered Search

Conduct Investigations

Our updated AI-powered search feature accepts a variety of natural language inputs so you can search your footage using everyday language.

Discover Insights

Whether its locating a person or tracking a vehicle, the search feature provides a unified solution for a diverse range of queries.

Check Your Results

Searches are subject to the limitations of the AI models used by the feature and other performance factors. Results may be inaccurate or incomplete in some cases. You should exercise discretion and use care when utilizing this feature.

AI-POWERED SEARCH TERMS

Our AI-powered search feature is designed to provide an enhanced and efficient search experience. However, search results are subject to the limitations of the AI models used and other performance factors, such as image resolution, camera placement, and lighting. The quality and consistency of the results may vary, particularly for searches based on gender or race. By enabling this feature, you understand that results may be inaccurate or incomplete in some cases, and therefore, should not be relied upon as the sole basis for decision-making. Results are provided 'as-is' without any warranty. You assume the risk of any incorrect or incomplete information and should review all results carefully before taking action.

You are responsible for your queries and should exercise discretion in the searches you perform. This feature may not be used for any illegal, harmful, harassing, or abusive activity, or in a way that violates any natural person's rights. You also agree not to circumvent, attempt to circumvent, or disable any protective mitigations implemented by Verkada.

All user queries and related metadata are logged and accessible to customer administrators for security and auditing purposes. By using this AI search feature, you acknowledge and accept these terms for responsible use. If you have concerns, please provide feedback in Command or contact our support team for assistance.

☑ I understand the Limitation and Responsible Use.

Conclusion

While our AI-powered search opens new possibilities for our customers, it's only the beginning. Our hybrid cloud architecture has, from the outset, enabled us to continue developing new CV capabilities and incorporate the latest technological advancements into our platform. Importantly, too, our hybrid cloud foundation affords us the flexibility to leverage a new foundation model should a more optimal one present itself, and also provides us with the tools to continuously implement new in-house enhancements. We look forward to offering our customers an easy, yet powerful and constantly-improving, search experience with privacy and respect top-of-mind.