Amazon Nova Multimodal Embeddings: Technical Report and Model Card

Amazon Artificial General Intelligence

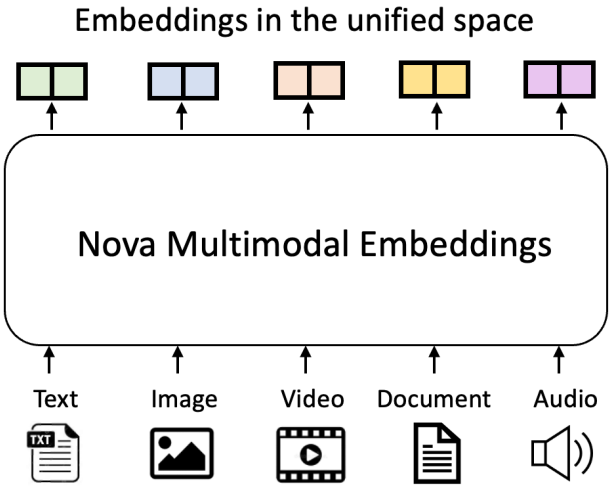

Figure 1: Amazon Nova Multimodal Embeddings model.

Abstract

We present Amazon Nova Multimodal Embeddings (MME), a state-of-the-art multimodal embedding model for agentic RAG and semantic search applications. Nova MME is the first embeddings model that supports five modalities as input: text, documents, images, video and audio, and transforms them into a single, unified embedding space. This powerful capability enables cross-modal retrieval — allowing users to search and find relevant information across different types of data. Nova MME supports up to 8K context length in the forms of text, images, video, documents and audio, converting them into numerical representations known as embeddings. These embeddings capture the semantic meaning of the underlying content, making it possible to compare, search, and perform reasoning tasks across modalities. By calculating the distance between embeddings, customers can power a wide range of intelligent applications — from semantic search and RAG-powered Large Language Models (LLMs) to content classification and beyond. It is the first unified embedding model that supports text, documents, images, video, and audio through a single model, giving developers the flexibility and performance needed to build next-generation AI solutions.

Enterprise API	Text	Image	Document	Video	Audio
Google Gemini Embedding [1]	✓				
Google Vertex Multimodal Embeddings [2]		✓		✓	
Amazon Titan Text Embeddings [3]	✓				
Amazon Titan MM Embeddings [4]		✓			
Cohere Embed 4 [5]	✓		✓		
TwelveLabs Marengo 2.7 [6]		✓		✓	✓
Amazon Nova MME	✓	✓	✓	✓	✓

Table 1: Use cases supported by existing embedding enterprise services. *Note*: TwelveLabs Marengo 2.7 supports embedding of text up to 77 tokens [7]. Given this limited context, use cases requiring long-form text embeddings (e.g., full-document similarity or large paragraph indexing) may not be optimally served by this model; its suitability should be validated by customers. Similarly, Google Vertex Multimodal Embeddings only supports up to 32 tokens. Given Cohere Embed 4's alignment to text and document use-cases, we opted to benchmark them accordingly [5].

1 Introduction

As industries evolve, more organizations are building advanced applications that harness multimodal semantic search and Retrieval-Augmented Generation (RAG) to unlock value from their vast stores of unstructured data. These repositories are rich and diverse — encompassing text, images, video, audio, and documents that frequently blend text and visuals in complex ways. Many businesses face unique challenges in accessing and making sense of this multimodal content. Examples include: (1) retrieving specific moments from video archives using rich, contextual queries, (2) searching financial documents that contain infographics along with text to identify the most relevant ones, and (3) grounding LLM outputs in highly contextual, multimodal data with higher fidelity. These use cases demand a robust solution that can understand and connect insights across different data types — something that traditional, single-modality embedding models do not provide. This is where multimodal embeddings (MME) become essential, offering a unified approach to process and retrieve information, delivering the depth and accuracy businesses need to derive actionable insights.

Despite the growing demand for multimodal capabilities, the current ecosystem remains highly fragmented. Most existing embedding models are tailored to only one or two modalities (as shown in Table 1), forcing organizations to maintain multiple disjointed embedding systems, with no straightforward way to align or map data across different content types, leading to isolated data silos. This fragmentation severely limits the ability to conduct unified searches or perform cross-modal reasoning across mixed-modality repositories. The problem becomes even more pronounced with complex, inherently multimodal content — such as documents combining text and infographics, or videos containing visual, audio, and temporal signals. Current models fail to effectively capture cross-modal relationships, underscoring the critical need for a unified representation space that can encode all content types consistently. Such a space would enable seamless retrieval, cross-modal analysis, and smooth integration into downstream AI pipelines. To address this challenge, we introduce **Amazon Nova MME** — *the first enterprise-grade solution* capable of embedding text, images, document images, video, and audio into a shared semantic space. This unified representation enables a wide range of applications, including: unimodal use cases (e.g., image-to-image retrieval), cross-modal scenarios (e.g., text-to-video retrieval), and multimodal queries (e.g., performing a text search with an accompanying reference image).

To build Nova MME, we developed a novel training paradigm that progressively enhances the model's ability to generate high-quality embeddings. This approach combines multiple loss functions — including next-token prediction and contrastive loss — to ensure robust performance across modalities. Recognizing the customer's need for higher accuracy without sacrificing efficiency, we incorporated Matryoshka Representation Learning [8]. This innovation allows Nova MME to support multiple embedding dimensions — specifically 3072, 1024, 384, and 256 — giving customers the flexibility to balance precision and performance. Smaller embeddings reduce storage requirements and speed up retrieval latency, making Nova MME a scalable and cost-effective solution for real-world deployment.

2 Evaluation

Retrieval performance benchmarking: We focused our evaluation on a range of cross-modal retrieval tasks. To assess the quality of Nova MME embeddings across different modalities, we measured the model's ability to accurately retrieve target items when queried using textual inputs. Specifically, we evaluated Nova MME on the following retrieval tasks: Text-to-Text, Text-to-Image, Text-to-Document, Text-to-Video and Text-to-Audio. For each task, we compare

	Amazon Nova Multimodal Embeddings	TwelveLabs Marengo 2.7	Google Vertex AI Multimodal Embeddings	Cohere Embed 4	Amazon Titan Multimodal Embeddings	Amazon Titan Text Embeddings V2	Google Gemini Embedding
VIDEO RETRIEVAL							
ActivityNet Avg (recall@1, recall@5, recall@10)	**81.2**	51.0	55.7	–	–	–	–
DiDeMo Avg (recall@1, recall@5, recall@10)	**80.6**	64.6	54.3	–	–	–	–
AUDIO RETRIEVAL							
AudioCaps Avg (recall@1, recall@5, recall@10)	**69.9**	55.9[1]	–	–	–	–	–
VISUAL DOCUMENT RETRIEVAL							
ViDoRe V2 NDCG@5	**58.7**	–	–	53.6	–	–	–
IMAGE RETRIEVAL							
TextCaps Avg (recall@1, recall@5, recall@10)	**88.9**	75.7[1]	*	–	84.7	–	–
MSCOCO Avg (recall@1, recall@5, recall@10)	76.7	**77.6**[1]	70.1	–	74.8	–	–
TEXT RETRIEVAL							
MTEB (Multilingual) NDCG@10	63.8	–	–	*	–	58.3	**67.7**[1]

Table 2: Nova MME performance comparison with selected available enterprise embedding models. The results for the publicly available models were obtained as follows: [1] indicates that the results are sourced from the model providers' own technical report/blog; "−" indicates that the corresponding modality use case is not supported; "*" indicates that benchmark results are not available as of 10/2025 to the Nova team; finally, all the remaining results were obtained by the Nova team. We evaluated Google Vertex AI Multimodal Embeddings using their official API and Amazon Titan, Cohere and TwelveLabs using the public Bedrock API, which offers all these models.

Nova MME with existing solutions and report the results in Table 2. Since no other single solution supports all the modalities that MME does, we elected to compare against multiple solutions in order to evaluate each modality.

Metrics: To assess the performance of Nova Multimodal Embeddings (MME) across various retrieval tasks, we use two primary metrics that align with standard practices in the evaluation of embedding models: (1) Normalized Discounted Cumulative Gain (NDCG) evaluates the quality of a ranked list by measuring how well the order of the results matches their relevance. NDCG@5 evaluates the top 5 ranking items in the list, while NDCG@10 the top 10; and (2) Recall@N evaluates whether the single most relevant content is included in the model's top N retrieved ones. For completeness, we compute an Average of recall at 1, 5 and 10.

Text-to-Text Retrieval: The text-to-text retrieval task aims to retrieve the most relevant text passage given a text query. We evaluate Nova MME retrieval performance on the Massive Multilingual Text Embedding Benchmark (MMTEB) (Multilingual, v2) [9] that contains content in 116 languages. Following the benchmark's standards, we measured performance in terms of NDCG@10. Overall, Nova MME offers competitive performance, surpassing Amazon's older model (Titan Text Embeddings v2) by 5.5pps.

Text-to-Image Retrieval: Evaluation on text-to-image retrieval probes the model capability in aligning detailed text captions with their corresponding visual concepts. We evaluate the retrieval performance on general images using MSCOCO [10] and on scene-texts using TextCaps [11]. Nova MME is more accurate then than Google Vertex Multimodal Embeddings and Amazon Titan Multimodal Embeddings by 6.6pps and 1.9pps in terms of the average recall, respectively.

Text-to-Document Retrieval: We evaluate text-to-document capabilities on the ViDoRe v2 benchmark [12]. This benchmark contains multilingual search queries (English, French, Spanish German) that closely simulate general, real world document retrieval use cases that have interleaved text and images. Compared to Cohere Embed 4 [5], Nova MME achieves 5.1pps higher average NDCG@5.

Text-to-Video Retrieval: We evaluate text-to-video performance on two of the most widely used video datasets: ActivityNet [13] and DiDeMo [14] against TwelveLabs Marengo 2.7 and Google Vertex Multimodal Embeddings. In both datasets, Nova MME achieves the highest average recall performance, surpassing them by +30.2/+25.5 on ActivityNet and +16/+26.3 points on DiDeMo, respectively.

Text-to-Audio Retrieval: We evaluate text-to-audio retrieval on AudioCaps [15] that contains natural language descriptions of audio streams in the wild. Among the enterprise solutions, TwelveLabs Marengo 2.7 is the only model we identified that supports audio modality and Nova MME outperforms it by 14pps.

Task	Benchmark	Embedding Size			
		3072	1024	384	256
Video Retrieval	ActivityNet	81.2	80.5	78.5	77.1
	DiDeMo	80.6	79.7	77.6	74.8
Audio Retrieval	AudioCaps	69.9	69.0	67.3	66.3
Visual Document Retrieval	ViDoRe v2	58.7	57.7	53.4	50.2
Image Retrieval	TextCaps	88.9	87.9	85.6	83.1
	MSCOCO	76.7	75.6	72.9	70.6
Text Retrieval	MTEB	63.8	62.9	60.4	58.3

Table 3: Nova MME performance with different embeddings sizes. Metrics are the same used in Table 2.

Retrieval performance of different embedding sizes: Nova MME employs a Matryoshka representation approach [8], which allows smaller embedding sizes to be derived directly from the larger ones—without requiring any additional computation. For instance, to facilitate rapid search for customers, Nova MME supports 256-dimensional embeddings, which can be obtained simply by selecting the first 256 values from the 3072-dimensional embeddings. This same principle applies to the 384- and 1024-dimensional embeddings, offering users a balance between search speed and accuracy. This design provides flexibility, enabling customers to generate the full, high-dimensional embeddings once and then choose the most suitable size for each specific application. Moreover, training with the Matryoshka representation helps maintain the richness of the learned representations and improves retrieval performance—even for the smaller sizes—ensuring strong performance across all dimensions. The effectiveness of this approach across different embedding sizes is demonstrated in Table 3.

Additional customer's guidelines: While we curated data from over 200 languages and achieved promising results on multilingual benchmarks (Table 2), we recommend customers perform their own testing to determine the utility of the model for their desired languages for text and document use cases. Similarly, we also recommend broader testing for audio use cases, as our benchmarking was limited in speech content recognition use cases.

3 Responsible AI

Our foundational approach to Responsible AI (RAI) for Nova MME is [16], structured around eight AWS RAI dimensions [17]. We defined the details of implementing these design objectives based on feedback from subject matter experts and applied them to the development of Nova MME. We focus on implementing appropriate guardrails during data collection, training, and deployment of runtime mitigation. Post the development of the system, we focus on rigorous evaluations using a collection of large in-house tests and red teamers. Below we outline certain processes we followed during the development of the MME model.

Data curation process: The data curation process for Nova MME was carefully designed with a strong focus on responsible AI principles. We sourced our data from trusted sources maintaining a lineage of data flow, and we implemented a multi-stage filtering pipeline to identify and remove content unsuitable for inclusion, such as suspected child sexual abuse material (CSAM). Such a data curation process ensures responsible data collection and processing and we also expect a positive impact on our downstream model evaluations.

Additional guardrails: Post building of the core MME model, we apply guardrails to adhere to RAI dimensions. These filters include filtering for known Child Sexual Abuse Monitoring (CSAM). We also implemented a fairness enhancement mechanism to ensure reduction of bias in the generated embeddings [18].

Evaluation: Our evaluation strategy for Nova MME comprised of offline evaluations and red teaming exercises. We built products such as image retrieval using Nova MME and evaluated models for bias and robustness against adversarial attacks. Our evaluations demonstrated reduction in bias post inclusion of fairness enhancement approaches and robustness to adversarial attacks. Additionally, we conducted rigorous red teaming in collaboration with our in-house experts and third party evaluators. We also performed testing to assess the robustness of MME against adversarial actors (who can perturb inputs/ indexed data) and developed mitigation including suggested guardrails to the customers.

References

[1] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. Gemini embedding: Generalizable embeddings from Gemini. *arXiv preprint arXiv:2503.07891*, 2025.

[2] Google Vertex Multimodal Embeddings. https://cloud.google.com/vertex-ai/generative-ai/docs/embeddings/get-multimodal-embeddings.

[3] Amazon Titan Text Embeddings v2 model. https://docs.aws.amazon.com/bedrock/latest/userguide/titan-embedding-models.html.

[4] Amazon Titan Multimodal Embeddings G1 model. https://docs.aws.amazon.com/bedrock/latest/userguide/titan-multiemb-models.html.

[5] Cohere Embed 4. https://cohere.com/blog/embed-4.

[6] Twelvelabs Marengo 2.7. https://www.twelvelabs.io/blog/introducing-marengo-2-7.

[7] Twelvelabs Marengo 2.7. https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-marengo.html.

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

[9] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. MMTEB: Massive Multilingual Text Embedding Benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in COntext. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[11] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: A dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer, 2020.

[12] Quentin Macé, António Loison, and Manuel Faysse. ViDoRe Benchmark V2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*, 2025.

[13] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[14] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[15] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. 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)*, pages 119–132, 2019.

[16] Amazon AGI. The Amazon Nova family of models: Technical report and model card. *Amazon Technical Reports*, 2024.

[17] Amazon. Building AI responsibly at AWS. https://aws.amazon.com/ai/responsible-ai/, 2024. Accessed: 2024-11-20.

[18] Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. Efficient fair PCA for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5250–5270. PMLR, 2023.