RESEARCH ARTICLE OPEN ACCESS

Architectural Challenges in Building Scalable B2B Products Using Multimodal AI Models

Shankar Krishnan
Product Manager, AWS; Boston, USA

Abstract:

The article presents a comprehensive analysis of the architectural challenges that arise when building scalable B2B products based on multimodal models. The study draws on a systematization of modern approaches to integrating textual, visual, voice, and structured data, as well as on an examination of architectures designed for extended context, sparse expert modules, adaptive layers, and composite agent-based systems. The discussion addresses how multimodality affects computational cost, interface stability, data organization, and corporate infrastructure requirements. Particular attention is given to the interplay between accuracy and scalability. It is shown that architectural choices create their own trade-offs, shaping both the depth of data interpretation and the degree to which the model depends on computational resources. The article analyzes key limitations that emerge in B2B environments, including data heterogeneity, the complexity of modality integration, preprocessing requirements, and the multi-layered nature of enterprise orchestration. The findings indicate that the performance of multimodal systems depends not only on model quality but also on the coherence of infrastructural, computational, and organizational components. The study concludes that multimodal AI becomes a system-forming element of corporate platforms, defining requirements for architecture, data, and engineering processes. The article will be useful for AI researchers, corporate product developers, B2B system architects, and specialists implementing multimodal solutions within large digital ecosystems.

Keywords: multimodal models, B2B architectures, scalability, modality integration, corporate platforms, architectural constraints, extended context.

Introduction

Multimodal models are becoming a pivotal element of B2B products, where data encompasses images, documents, voice, video, and telemetry. Competitive advantage is gained by solutions capable of integrating these sources into a unified semantic environment. Consequently, architectures combining visual encoders, language models, and adaptation modules are acquiring increasing significance.

The effectiveness of corporate systems is determined by how accurately a model interprets real-world objects and processes. Approaches utilizing adaptation layers, sparse experts, and advanced normalization mechanisms allow for reduced token costs, faster modality alignment, and increased noise resilience—factors critical for B2B systems with high requirements for latency and stability.

Multimodal analytics is evolving into a business metric in tasks involving documentation, dialogue analytics, and autonomous systems; accuracy depends directly on architectural decisions. In applied fields such as supply chains, corporate services, or automotive platforms, engineering optimizations exert a notable influence on the reliability and quality of inferences. In the automotive industry, this is particularly crucial due to the link between the interpretation of visual and voice data and safety levels.

Different B2B scenarios require distinct architectures. Client services necessitate speed, documents require precise recognition, and automotive systems demand determinism and resource economy. A universal multimodal stack does not exist. Successful solutions are built around adapting models to the specifics of corporate data and integrating them into existing processes. The scientific novelty lies in systematizing the architectural factors that determine the scalability of multimodal AI in B2B products and identifying the dependencies between model type, data quality, and service stability.

The objective of the study is to define the architectural challenges arising in the creation of scalable B2B products using multimodal AI models and to identify mechanisms that ensure a balance between accuracy, speed, and stability. To achieve this goal, the following tasks were addressed: an analysis of modern multimodal architectures was conducted, corporate infrastructure limitations were defined, approaches to fusion, normalization, and orchestration were compared, and recommendations were formed for integrating models into complex B2B systems.

The research hypothesis posits that the scalability of B2B products depends directly on the coordinated operation of modular multimodal architectures: the stability of fusion, the effectiveness of adaptation layers, and the quality of computational process orchestration. Their combined application reduces data processing uncertainty, lowers operational costs, and increases the accuracy of interpreting corporate scenarios.

The scope of the study is limited to the sphere of corporate B2B solutions, where multimodal artificial intelligence is used in document processing, comprehensive analytics, voice interfaces, business services, and automotive systems. Production, logistical, and business processes are considered only as a context defining the requirements for scalability, latency, security, and model integration into existing platforms.

Materials and Methods

The methodological foundation of the study is formed at the intersection of architectural approaches to building multimodal systems, data engineering, and corporate B2B product development practices. This interdisciplinary approach allows for the unification of several levels of analysis: enterprise infrastructure limitations, the specifics of multimodal data processing, and the architectural solutions of modern models applied in a corporate environment. Source selection was conducted based on the criteria of scientific reliability and relevance. The analysis included works from 2022 to 2025 published in peer-reviewed journals and presented in open repositories.

The study by Allec et al. [1] demonstrates that multimodal and multi-institutional data require strict standardization and complex aggregation, setting the framework for corporate system architectures. Gerling et al. [2] emphasize the necessity of combining visual and textual representations in document analytics, which requires unified models. The Granite Vision Team [3] demonstrates the effectiveness of lightweight multimodal models with optimized encoders for corporate tasks.

Huang et al. [4] point to the importance of unified masking of text and images in models for documents. Kandogan et al. [5] describe the principles of compound AI systems with data streams and agent registries. Li et al. [6] show that BLIP-2 reduces computational costs by combining frozen visual encoders with language models.

Li et al. [7] confirm the effectiveness of Mixture-of-Experts architectures for combining modalities. Ruan [8] emphasizes the significance of correct multimodal fusion for increasing the accuracy of industry solutions. Shakhadri et al. [9] demonstrate the role of QK-normalization and hybrid training schemes in improving the efficiency of corporate models. Srinivas et al. [10] highlight the importance of the coordinated operation of agentic multimodal systems in B2B scenarios.

The methodological strategy of the research is based on a systematic analysis of architectural solutions, data standardization procedures, and engineering methods described in the cited sources. The synthesis of the obtained data allowed for the identification of key directions—data standardization, modality adaptation models, and agentic component orchestration systems—as the defining parameters of scalability for B2B solutions based on multimodal AI.

Results

The scalability of multimodal systems in corporate products is determined not by the diversity of modalities but by the extent to which the architecture allows for managing data flow complexity and interpretation accuracy. In a B2B product environment, a model encounters documents, schematics, interfaces, visual dashboards, and telemetry, and each of these modalities imposes its own requirements for contextualization depth. The analysis results indicate that architectural decisions become the key factor determining product stability and extensibility.

One of the most illustrative directions is the principle of unified document representation described in the study by Huang et al. [4]. Unified masking of images and text sets a general contour for data processing, facilitating further scaling. In corporate conditions, this implies a reduction in the volume of custom components and simplified model maintenance. However, such an approach limits flexibility. Universality increases modality compatibility but reduces processing variability.

A different trajectory is observed in the architecture of compound systems proposed by Kandogan et al. [5]. Data streams, schedulers, and agentic contours form an environment in which the model becomes only one element of the computational process. Here, data interpretation is distributed among agents, and multimodality manifests not at the model level, but at the system level. For B2B products, this changes the very principle of design. Instead of a "strong model," "strong orchestration" is created. The adaptation logic demonstrated by Li et al. [6] is also of interest. Using an intermediate transformer between the visual encoder and the language model turns the model into a construction set where components can be replaced without disrupting the overall process. This approach demonstrates that scalability can result from modularity rather than simply increasing computational resources.

Sparse expert subsystems, developed by Li et al. [7], show that multimodality can be selective. Each modality is serviced only by those experts for whom it is relevant. This alters the approach to scaling itself. The model grows not horizontally, but fragmentally, as new data types appear. Optimization techniques applied by Shakhadri et al. [9] demonstrate yet another path. Architecture is used as a means to reduce data volume requirements. QK-normalization and hybrid normalization schemes do not expand the model but make it robust to small and noisy samples. In B2B environments, where data is often limited by availability and confidentiality, this transforms into a strategic advantage. Table 1 presents a comparison of the key architectures that formed the basis of the analysis.

Table 1 – Comparison of Multimodal Model Architectures (Compiled by the author based on sources: [4, 5, 7, 9])

Model / Architecture	**Key Components**	**Fusion Type**	**Specific Features**
LayoutLMv3	Unified masking, Transformer	Feature-level fusion	Single pipeline for text–image processing
Compound AI	Agents, streams, planners	Orchestration fusion	DAG planners, streaming, agent workflow
BLIP-2	Frozen ViT, Q-Former, frozen LLM	Late fusion	Q-Former as ViT→LLM adaptor
Uni-MoE	Sparse MoE	Expert fusion	Sparse experts and modality connectors
Shakti-VLM	QK-Norm, hybrid norm, RoPE-2D	Embedding fusion	Three-stage training, dynamic resolution

The comparison of architectures demonstrates that multimodality in B2B products is not a singular technical approach. Some solutions strive for unification, minimizing processing variability, while others aim for modularity or selectivity. Each of these strategies forms its own scalability trajectory. The analysis reveals that successful corporate systems align not with a universal architecture, but with the alignment of the architectural principle with the nature of the data: document products benefit from unification, industrial analytics systems from expert sparsity, and high-load services from orchestration. It is this ability of the architecture to "adjust" the model to the data that determines the stability of B2B multimodal solutions.

The analysis of modern multimodal solution architectures shows that their scaling in a corporate environment faces model quality issues and fundamental infrastructure limitations. These limitations manifest at the level of data, computational processes, modality fusion mechanisms, and workflow pipeline organization. The aggregate of factors indicates that the stability of multimodal systems is determined by the neural network architecture and the architecture of corporate computing as a whole. Table 2 examines the systemic distribution of key limitations identified during the source analysis.

Table 2 – Architectural and infrastructure limitations of multimodal systems (Compiled by the author based on sources: [1, 5, 10])

Category	Description
Data heterogeneity	Heterogeneous data and complex preprocessing
Fusion cost	High computational cost of multimodal fusion
Latency constraints	Strict SLAs in B2B products
Alignment difficulty	Difficulty of text–image alignment
Dependency on data quality	Model effectiveness increases only with high-quality data
Orchestration overhead	Complex agentic pipelines

The identified limitations demonstrate that multimodal AI technology in the corporate segment develops under conflicting requirements. On one hand, organizations strive to use complex visual, textual, and structured data to improve decision accuracy. On the other hand, the very nature of such data sets sets a high entry barrier. The study by Ruan [8] shows that multiformat and noisy corporate data require expensive normalization and careful alignment before they can be included in analytical pipelines.

The cost of combining modalities is of particular importance. Shakhadri et al. [9] emphasize that even with optimized mechanisms—such as QK-normalization and hybrid normalization schemes—multimodal fusion remains a computationally intensive process. The analytical conclusion here is that architectural optimizations do not eliminate the problem entirely but merely shift it toward more efficient resource allocation. This implies that the scalability of such systems depends not on the "power" of the model, but on the architecture's ability to correctly manage data streams and their priorities.

In corporate products, the response time factor dominates. Kandogan et al. [5] detail the architecture of compound systems, where task schedulers, agent pipelines, and streaming mechanisms create additional latency. Under the strict SLAs characteristic of the B2B segment, even small latency fluctuations lead to business process failures. A structural compromise arises: the more intelligent the system, the harder it is to ensure stable response times.

The issue of data quality requires separate attention. Ruan [8] shows that models demonstrate significant accuracy growth only with high-quality input data. This leads to an important practical conclusion. Increasing model effectiveness is achieved by perfecting the architecture and optimizing corporate processes for data collection, standardization, and labeling. Otherwise, the system remains limited by the "physics" of its data, regardless of how advanced its algorithms are. Finally, agentic orchestration, acting as a promising approach, itself creates new sources of complexity. In the architecture described by Kandogan et al. [5], multi-level interactions of agents, streams, and schedulers increase operational costs. A self-complication effect emerges. The more flexible the system becomes, the harder it is to control its behavior in real-world conditions where data flows are unpredictable and contexts change dynamically. The BLIP-2 method [6] shows that using frozen visual encoders and language models in combination with a lightweight Q-Former adapter allows for maintaining high performance with a reduced number of trainable parameters.

Thus, the infrastructure limitations of multimodal systems form a complex of barriers that cannot be eliminated solely by architectural methods. Scalability is determined by data quality, pipeline stability, organizational structure predictability, and the accuracy of modality fusion mechanisms. In a corporate context, these factors form an efficiency limit that must be considered when designing B2B solutions based on multimodal AI.

Discussion

The evolution of multimodal models observed in studies shows that their capabilities significantly outstrip the capacities of the corporate infrastructures into which these systems are intended to be embedded. Therefore, a key feature of real-world implementation becomes not the attempt to reproduce research results, but the necessity to align model requirements with corporate environment limitations.

The first contradiction is related to context volume. In the work of Shakhadri et al. [9], Shakti models operate with sequences of 16–32 thousand tokens. However, in B2B product conditions, such context volumes conflict with habitual resource norms. Corporate systems are initially designed for predictable loads, not for sharp spikes in memory consumption. As a result, expanded context turns into a factor that limits scalability rather than strengthening the model, because the infrastructure is forced to adjust to the model, not vice versa.

The next source of contradictions involves modality fusion mechanisms. The late fusion architecture described by Li et al. [6] and the use of sparse experts in the work of Li et al. [7] create a specific effect. The more flexible the model becomes, the more rigid the requirements it imposes on the computational pipeline. Multimodal fusion proves to be not just a model step, but a center of technological tension. In corporate products where predictable latencies are important, such pipelines disrupt the balance between system intelligence and operational stability. A no less significant contradiction arises at the level of preliminary data processing. LayoutLMv3, as shown in the study by Huang et al. [4], initially assumes a strictly structured input. Meanwhile, the multimodal analysis of supply chains presented in the study by Ruan [8] demonstrates the dependence of accuracy on data quality. These observations collectively indicate that the effect of multimodality is limited not by model architecture, but by data preparation discipline. Corporate ecosystems rarely ensure stable input flow quality, creating a fundamental divergence between ideal and real functioning conditions.

The most notable contradiction is traced in the organization of computations. The architecture of streams, agents, and schedulers described by Kandogan et al. [5] allows for building flexible compound systems, but this very flexibility creates a "self-complication" effect. A system calculated for adaptive behavior begins to depend on constant synchronization between components. Consequently, a corporate product is forced to simultaneously ensure intelligence and stability, yet these two requirements are poorly compatible when working with multimodal models.

Therefore, the main difficulty in implementing multimodal models lies not in the models themselves, but in the mismatch between their operational logic and the logic of corporate ecosystems. Contradictions manifest in context size, fusion complexity, dependence on data quality, and the multi-level organization of processing. It is these divergences that determine scalability limits and the conditions under which multimodal AI can be included in B2B architectures without a loss of stability.

The question of balance between computational cost and accuracy of multimodal models remains key for corporate systems. Various architectural solutions form their own trade-off trajectories, defining which characteristics—speed, scalability, interpretation depth, or economy—can be prioritized. Table 3 shows how the architectures presented in the sources realize this balance.

Table 3 – Trade-offs between performance and scalability in multimodal models (Compiled by the author based on sources: [2, 4, 7])

Model	Trade-off	Basis
BLIP-2	Minimal training → high efficiency but limited context	Q-Former + frozen ViT/LLM
Uni-MoE	Scalability via sparse experts but complex routing	Sparse experts, modality connectors
Shakti -VLM	High accuracy with fewer tokens but complex 3-stage training	QK-Norm + hybrid norm + 487B tokens

The architectures under review demonstrate that the trade-off between accuracy and scalability cannot be viewed as a secondary consequence of model selection. On the contrary, it becomes a fundamental property of the multimodal approach. Each decision regarding modality fusion, context management, or training optimization forms its own constraint, determining how much the model can be adapted for the corporate environment.

The BLIP-2 architecture proposed by Li et al. [6] represents the most economical option. The use of frozen visual encoders and a language model reduces computational expenses and allows for system implementation with limited resources. However, this economy is achieved at the price of simplified context representation. Late modality fusion proves effective only when the data structure is predictable and scenarios are limited. In corporate systems, where data is uneven and variable, such a model demonstrates stability only in narrow segments. The Uni-MoE architecture described by Li et al. [7] offers an opposing strategy. Scaling is achieved through sparse experts, allowing for dynamic load distribution and the connection of specialized modules for specific modalities. However, this flexibility creates its own limit. Routing between experts becomes more complex as the number of modalities increases. In corporate products, such an approach requires a highly stable infrastructure; otherwise, scalability growth is accompanied by a drop in response predictability.

Shakti-VLM models, presented by Shakhadri et al. [9], offer a different type of trade-off. Increased accuracy is achieved through architectural optimizations—QK-normalization, hybrid normalization, and encoders with improved stabilization. At the same time, the volume of tokens used reduces the model's dependence on training set sizes. However, such a training scheme requires three independent stages and complex parameter alignment, limiting implementation speed and complicating system maintenance. In corporate use, this means the gain in accuracy is achieved at the cost of complicating the engineering cycle.

These examples confirm that the trade-off between accuracy and scalability arises not as a consequence of hardware limitations, but as a regularity determined by the very nature of multimodal architectures. Models focused on economy lose in universality; models striving for flexibility face rising organizational complexity; systems achieving high accuracy require disproportionately complex training.

Thus, the analysis of trade-offs shows that the scalability and accuracy of multimodal models exist in a relationship of mutual limitation. Architectural choice determines model performance and its capacity to be integrated into corporate solutions without a loss of stability. These trade-offs form the practical framework within which multimodal AI for B2B systems develops.

Conclusion

The study conducted has shown that the architectural efficiency of multimodal models in B2B systems is determined not by individual elements—context, modality fusion mechanisms, or training schemes—but by the coherence of the entire computational and organizational structure into which these models are built. Multimodal AI becomes effective only when data architecture, execution infrastructure, and orchestration mechanisms form a unified technological contour.

The analysis revealed that the key factor in successful scaling is the system's ability to work with heterogeneous data streams while maintaining stable latencies and predictable quality. Increasing context, complicating fusion schemes, and rising requirements for data preparation inevitably intensify the load on corporate infrastructure. Therefore, the architectural solution turns out to be not merely a means of increasing accuracy, but a tool for managing trade-offs between computational cost, stability, and implementation flexibility.

It is demonstrated that multimodal models influence the product cycle as algorithmic modules and elements that change the structure of the B2B application itself. They require restructuring service interaction logic, modernizing processing pipelines, and clarifying responsibility boundaries between components. Thus, the implementation of multimodal AI becomes a systemic task where engineering, organizational, and product decisions prove to be interconnected.

The results obtained form the understanding that the efficiency of multimodal systems is determined by a combination of three factors: architectural coherence, resilience to data variability, and the infrastructure's capacity to support complex models without a loss of reliability. It is these parameters that set the limits for realizing multimodal technologies in B2B products and outline directions requiring priority development.

The conducted analysis supports the assertion that multimodal architectures possess significant potential in corporate B2B solutions, yet their practical implementation requires thoughtful management of computational resources, operational reliability, and observability mechanisms. Promising directions for further research include: the development of universal benchmarks for evaluating multimodal systems in production scenarios; the study of routing stability in MoE architectures; and the creation of quality monitoring standards for multimodal models.

Thus, multimodal AI in the corporate environment acts not as a separate technology, but as a principle of building digital solutions in which architecture, data, and computation are viewed as a unified system. This creates a foundation for future research oriented toward developing stable platforms capable of supporting increasing model complexity without reducing manageability, transparency, and operational predictability.

References

1. Allec, S. I., Muckley, E. S., Johnson, N. S., & others. (2024). A case study of multimodal, multi-institutional data management for the combinatorial materials science community. Integrating Materials and Manufacturing Innovation, 13, 406–419. https://doi.org/10.1007/s40192-024-00345-7

2. Gerling, C., & Lessmann, S. (2023). Multimodal document analytics for banking process automation. arXiv. https://doi.org/10.48550/arXiv.2307.11845

3. Granite Vision Team. (2025). Granite Vision: A lightweight, open-source multimodal model for enterprise intelligence. arXiv. https://doi.org/10.48550/arXiv.2502.09927

4. Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022). LayoutLMv3: Pre-training for Document AI with unified text and image masking. arXiv. https://doi.org/10.48550/arXiv.2204.08387

5. Kandogan, E., Rahman, S., Bhutani, N., Zhang, D., Chen, R. L., Mitra, K., Gurajada, S., Pezeshkpour, P., Iso, H., Feng, Y., Kim, H., Shen, C., Wang, J., & Hruschka, E. (2024). A blueprint architecture of compound AI systems for enterprise. arXiv. https://doi.org/10.48550/arXiv.2406.00584

6. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv. https://doi.org/10.48550/arXiv.2301.12597

7. Li, Y., Jiang, S., Hu, B., Wang, L., Zhong, W., Luo, W., Ma, L., & Zhang, M. (2024). Uni-MoE: Scaling unified multimodal LLMs with a mixture of experts. arXiv. https://doi.org/10.48550/arXiv.2405.11273

8. Ruan, M. (2024). The application of multimodal AI large model in the green supply chain of the energy industry. Energy Informatics, 7, 98. https://doi.org/10.1186/s42162-024-00402-7

9. Shakhadri, S. A. G., Kr, K., & Angadi, K. B. (2025). Shakti-VLMs: Scalable vision-language models for enterprise AI. arXiv. https://doi.org/10.48550/arXiv.2502.17092

10. Srinivas, S. S., Das, A., Gupta, S., & Runkana, V. (2025). Agentic multimodal AI for hyperpersonalized B2B and B2C advertising in competitive markets: An AI-driven competitive advertising framework. arXiv. https://doi.org/10.48550/arXiv.2504.00338