GENERATIVE SYNCHRONOUS DIFFUSION · GSD

The Next Breakthrough in Multimodal AI

For Creators, Agents and Enterprises

A topological, graph theoretic architecture that collapses the multimodal scaling wall from $O(M^2N^2)$ to $O(M \log N + N)$, generating six synchronised, independently editable modalities in a single one shot pass with positive unit economics from day one.

WHAT THIS PAPER ESTABLISHES

85 to 90%	≈67×	99.89%	$2.00
Training cost reduction	Realised inference efficiency gain	Prompt constraint adherence	Per 60s, full multimodal

Animikh Roy

CEO, CTO & Product Architect
University of Sussex, Astrophysics & CS

Neel Roy
Co-Author, Head of Growth · Stanford University, CS

Public Technical Overview · May 2026
Wishtales AI Inc. · Delaware C-Corporation · Palo Alto, California
India subsidiary
PUBLIC EDITION · CLEARED FOR GENERAL DISTRIBUTION

Abstract

Multimodal artificial intelligence faces a structural paradox: capability advances rapidly while the cost of producing rich, synchronised output rises faster still. Humans perceive the world as simultaneous vision, sound and language, so systems that generate content must match that simultaneity to be useful.

This paper presents **Generative Synchronous Diffusion (GSD)**, an architectural paradigm that reframes multimodal synthesis as traversal and optimisation over a topological graph embedded in an n dimensional tensor space. GSD replaces the polynomial scaling of cross-attention with log-linear scaling. Across six synchronised modalities it reduces the dominant per generation cost from $O(M^2N^2)$, which is quadratic in both the modality count M and the sequence length N, to $O(M \log N + N)$. On the evaluation harness this corresponds to an approximately 67× realised inference efficiency gain at fixed quality, with 99.89% prompt constraint adherence and real world physics coherently represented across all modes. For users this means one shot generation that replaces a multi tool workflow while delivering an order of magnitude more value per dollar than multi pass alternatives. The conclusion is that sustainable multimodal deployment requires a change of computational complexity class, not incremental optimisation of the existing one. This public overview presents the result and its implications; the proprietary internals are intentionally not disclosed here.

> **Genesis of GSD technology**
>
> GSD draws conceptual inspiration from "A Novel Approach to Topological Graph Theory with R-K Diagrams and Gravitational Wave Analysis" (Roy & Kesselman, 2022; arXiv:2201.06923, Harvard ADS). That research is an inspiration, not a foundation: only the R-K toolkit, specifically the idea of representing complex relationships as R-K diagrams and the use of combinatorial optimisation over those structures, served as a conceptual starting point. GSD is an independent architecture; its design, optimiser and signature construction are proprietary and are not described in this public edition.

Contents

About this edition

This is the public technical overview. It states the architectural result, the complexity argument and the measured outcomes, and it explains why GSD is a paradigm shift, while deliberately omitting the proprietary mechanism. Figures are directional and conservative.

01 Introduction: The Economics of Multimodal Intelligence

Artificial intelligence has reached an inflection point at which the sophistication of multimodal capability collides with the realities of computational economics.

Third party industry projections place the broader multimodal AI market at roughly USD 13 billion in 2025, growing toward the high hundreds of billions of dollars by the mid 2030s at a compound annual growth rate in the mid forties of per cent. Within that, one widely cited estimate puts the AI video segment at USD 3.86 billion in 2024 rising to USD 42.29 billion by 2033, a compound annual growth rate of approximately 32 per cent measured over the 2025 to 2033 window (Grand View Research). These figures are external projections, directional only. The point that matters is structural: major providers allocate large sums to multimodal development, yet frontier video systems operate at materially negative gross margins because the unit cost of generation exceeds the price the market will bear.

This is not a pricing problem. It is an architectural one. Cross-attention computes pairwise interactions between every pair of elements across modalities, so the cost of synchronising M modalities over sequence length N scales quadratically in N and quadratically in M, that is, as $O(M^2N^2)$. For a system that must synchronise six modalities simultaneously this scaling makes large scale deployment economically prohibitive, which is why production cross-attention systems remain effectively limited to two synchronised modalities.

GSD reframes multimodal synchronisation through graph theoretic principles. By embedding modal relationships in an explicit topological structure and resolving synchronisation through structure rather than dense attention, GSD attains log-linear scaling while preserving semantic and physical coherence, removing the cause of the negative margin rather than subsidising its symptom.

$O(M^2N^2)$	$O(M \log N + N)$	$M = 2 \rightarrow 6$	26%+
The scaling wall competitors face	What GSD achieves	Synchronised modalities, in production	Blended gross margin, day one

> **Why this is an architecture problem, not a pricing problem**
>
> Across the industry, video systems operate at negative gross margins because dense cross-attention with quadratic scaling makes every additional synchronised modality and every additional second of output disproportionately expensive. Subsidising each generation is therefore a structural consequence of the chosen architecture, not a discount strategy. GSD changes the complexity class, which changes the cost structure at its source.

02 Theoretical Foundation and the Scaling Wall

The organising idea behind GSD is simple to state: complex relationships between signals can be encoded as stable structure in a high dimensional space and then composed, rather than recomputed from scratch for every pair of elements. This section defines the few variables needed to understand the result and then shows precisely where the industry wall comes from.

2.1 Defining the variables

Asymptotic notation follows standard usage: $O(\cdot)$ is an upper bound and $\Theta(\cdot)$ a tight bound. Only four quantities are needed to follow the argument.

Symbol	Definition
M	Number of synchronised modalities. Production cross-attention systems are effectively limited to M = 2; GSD operates at M = 6 (video, voice, music, sound effects, on screen text, and a graph or numerical reasoning track).
N	Sequence length, that is, the number of text tokens or image and video patches, typically between 77 and a few thousand.
L	Number of iterative denoising steps in a conventional diffusion model, commonly in the hundreds.
d	Embedding dimension of the latent representation; treated as a fixed hardware level constant in the asymptotic statements.

2.2 The $O(M^2N^2)$ multimodal wall

Cross-attention computes pairwise interactions between all elements across modalities. Synchronising M modalities requires attention over every unordered pair of modalities, of which there are $M(M-1)/2$. With each modality carrying a sequence of length N and embedding dimension d, a single pairwise cross-attention costs $\Theta(N^2d)$. Summing over all pairs and over the L denoising steps of an iterative diffusion model gives the total per generation cost:

$$C_{attn} = L \cdot [M(M-1)/2] \cdot N^2 d = \theta(L \cdot M^2 \cdot N^2 \cdot d)$$

Holding d and L as constants, the dominant growth is **$\Theta(M^2N^2)$**: quadratic in the modality count and quadratic in the sequence length. This single fact is why no production system exceeds two tightly synchronised modalities; the cost of a third, fourth or fifth modality grows faster than any pricing strategy can absorb.

2.3 Evolution beyond Mapper: from summary to semantics

Topological data analysis was advanced significantly by the Mapper algorithm of Singh, Mémoli and Carlsson (2007). Mapper summarises the shape of high dimensional data as a simplified graph and is best understood as a visualisation and exploration tool rather than a complete world model. GSD takes the spirit of that idea much further: instead of a partial summary it builds a complete hierarchical knowledge graph centred on Event-Nodes that act as static contextual hubs, with features clustered topologically around those events. The result is an explicit internal semantic world model rather than a summary, and it is this explicitness that makes the scaling result possible.

2.4 Graph embedding, in one idea

GSD reformulates synchronisation by constructing a unified graph whose vertices represent semantic units across all modalities and whose edges encode their relationships. Because semantically similar vertices are organised into a balanced hierarchical structure, locating one modality signature is a balanced tree search rather than an exhaustive comparison. The M modality signatures are retrieved independently and in parallel, so the retrieval cost is $O(M \log N)$ instead of the $O(M^2N^2)$ of dense attention. The detailed construction is proprietary; what matters publicly is the change of complexity class it produces.

2.5 GSD versus learned sparse attention

Recent efficient attention work, such as Moonshot AI's Mixture of Block Attention (MoBA), partitions context into blocks and applies a learned gate so each query attends only to a sparse subset, reducing complexity below quadratic. The relationships remain implicit in the trained weights and the sparsity is a statistical artefact of the gate. GSD takes an inherent structure stance: the topological graph is explicit and deterministic, and sparsity follows from the ontology rather than from learned routing.

Property	Standard Transformer	MoBA (Moonshot AI)	GSD Topological Transformer
Attention mechanism	Full (dense) attention	Block sparse attention	**Hierarchical graph traversal**
Sparsity method	None (dense)	Learned statistical gating	**Deterministic structural (ontological graph)**
World model	Implicit in weights	Implicit in weights	**Explicit topological graph**
End to end pipeline	$\Theta(L \cdot M^2 \cdot N^2)$	Sub quadratic in N	**$O(M \log N + N)$**
Inductive bias	Weak (positional)	Weak (limited structure)	**Strong (topological invariants)**

2.6 How GSD judges its own output

Every generated result is scored on two independent axes: whether it has the correct structure, the right things in the right relationships, and whether it is statistically plausible, that is, consistent with how the real world actually looks and sounds. A result is only accepted when it scores well on both. Conventional pipelines optimise primarily for the second axis and only approximate the first, which is why they drift, lose characters and break continuity over long sequences. The exact scoring function is proprietary; the diagram below shows the principle.

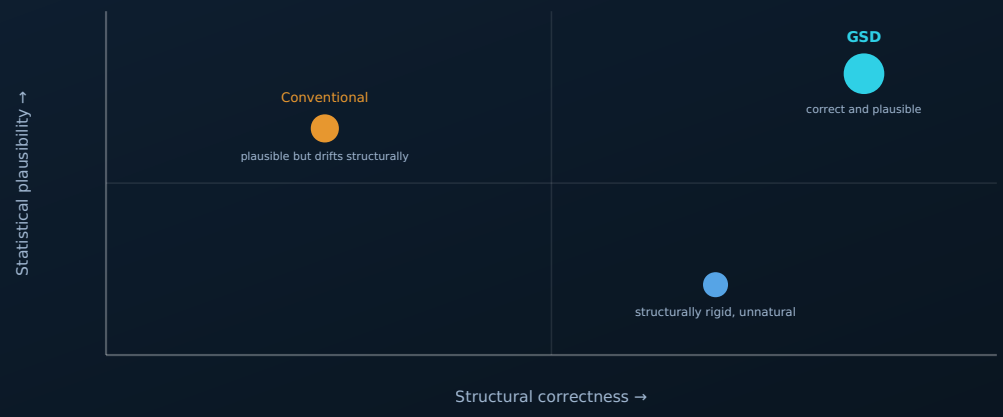

Figure 1. Quality is judged on two independent axes. GSD targets the upper right: structurally correct and statistically plausible at the same time, which is where consistency over long sequences comes from.

2.7 Numerical validation of the complexity class

The chart plots theoretical operation counts on a logarithmic axis for M = 6 modalities as the sequence length grows. Cross-attention follows $\Theta(M^2N^2)$; iterative diffusion multiplies this by L denoising steps to give $\Theta(L \cdot M^2 \cdot N^2)$; GSD follows $O(M \log N + N)$. The separation is not incremental, it is a change of complexity class, which is the entire point.

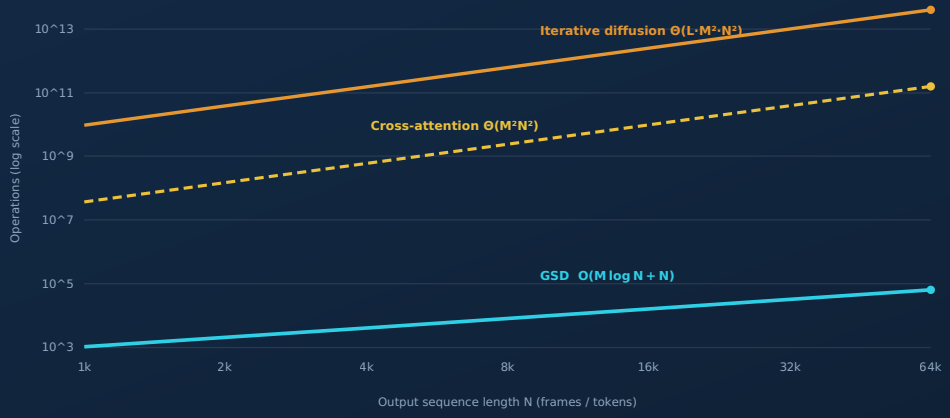

Figure 2. Theoretical operation count versus output sequence length. Cross-attention and iterative diffusion scale polynomially; GSD scales log-linearly.

N (sequence)	Cross-attn $\Theta(M^2N^2)$	Diffusion $\Theta(L \cdot M^2 \cdot N^2)$	GSD $O(M \log N + N)$	Ratio vs diffusion
1,000	3.60×10^{7}	9.00×10^{9}	**1.06×10^{3}**	8.49×10^{6}
4,000	5.76×10^{8}	1.44×10^{11}	**4.07×10^{3}**	3.54×10^{7}
16,000	9.22×10^{9}	2.30×10^{12}	**1.61×10^{4}**	1.43×10^{8}
64,000	1.47×10^{11}	3.69×10^{13}	**6.41×10^{4}**	5.75×10^{8}

Illustrative operation counts use M = 6, L = 250 and a log base of 2, with d and lower order constants absorbed into the asymptotic notation. The conclusion is independent of these constants: GSD converts a polynomial regime into a log-linear one, so the advantage widens monotonically with both sequence length and modality count. Realised hardware efficiency is lower than the theoretical bound and is reported conservatively in Section 4.

03 Architecture: A Paradigm Shift in How Synthesis Works

Conventional systems generate, then fix. GSD plans once, then renders once. That single inversion is the paradigm shift.

3.1 The Retrieve, Optimise, Render pipeline

GSD separates the problem into three stages that each do one job well. The retrieval stage finds the right structured concepts; the optimisation stage resolves all synchronisation in a single planning step; the render stage produces every modality from one shared plan. The internal workings of each stage are proprietary, but the shape of the pipeline is what gives the result.

RETRIEVE 01

Symbolic layer

O(M log N)

Finds the M candidate signatures by parallel search of the topological graph. No neural computation. Analogous to the prefrontal cortex retrieving structured concepts.

OPTIMISE 02

Bridge layer

O(1) amortised

Composes one geometrically consistent plan from the candidates. All synchronisation is resolved here. Analogous to the thalamus binding signals into one unified percept.

RENDER 03

Neural layer

O(N)

Renders the pre synchronised plan in a single linear pass. Analogous to the motor cortex executing a pre planned movement.

3.2 What changes, in one picture

Conventional production stacks generate video first, then synthesise and align audio, music and effects in separate passes and separate tools, accumulating error and cost at every handoff. GSD produces all six layers from a single pass over one shared plan, so there is nothing to re align afterward.

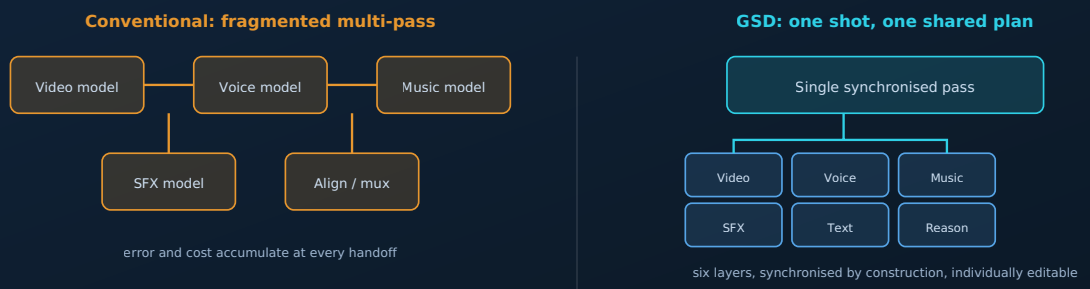

Figure 3. The architectural inversion. Conventional pipelines align after generation; GSD synchronises before it, so the six output layers cannot drift relative to each other.

Synchronisation by construction

Because every modality is produced from one shared plan rather than stitched together afterward, lip movement, footsteps, ambience and transitions are co produced and cannot fall out of sync. There is no separate lip sync model and no post hoc alignment step to fail.

3.3 One-shot versus multi-pass generation

Most current systems are multi pass: video is generated first, then audio is synthesised and mixed separately, which introduces alignment error and cumulative quality loss. Recent frontier systems do produce a synchronised audio track alongside video, but it is a single mixed track rather than separable layers. GSD generates all modalities from one pass, so each of the six layers remains independently editable.

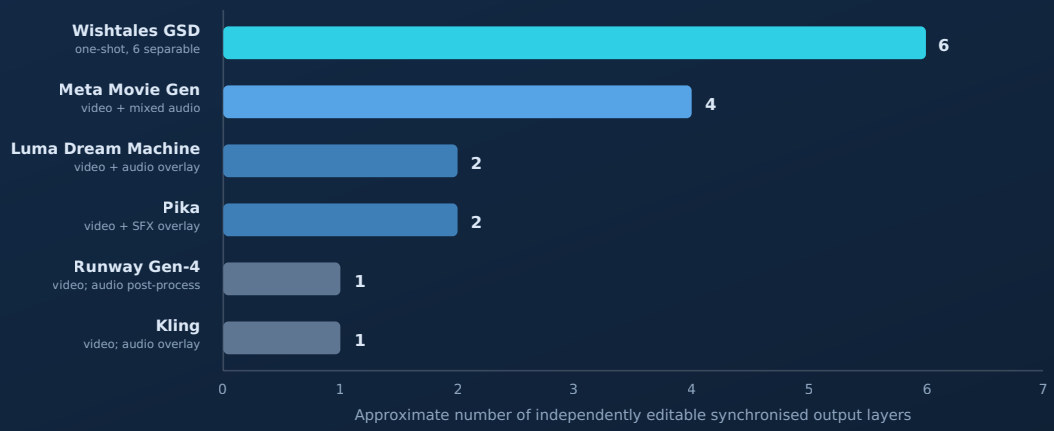

Figure 4. Independently editable synchronised output layers, an approximate characterisation based on publicly described capabilities. Competitor audio is typically a single mixed track rather than separable layers.

Among publicly described systems, Meta Movie Gen is the closest research preview to broad multimodal generation but is not generally available. Luma and Pika generate video first and overlay audio separately. Runway and Kling deliver video with post processed or overlaid audio at short clip lengths. GSD is distinguished less by the presence of audio than by producing six synchronised layers that remain separable and individually editable, at custom length, in a single pass.

Editing without re-rendering: Giga Papaya

Because the six layers are separable, the Giga Papaya capability lets a user change one layer, for example re voice a character or swap the music, without regenerating the whole asset. Conventional systems must re render the entire video for any change. This is a structural advantage of the one shot architecture; the mechanism is proprietary and is not described here.

04 Performance: Efficiency and Quality

Evaluation across 10,000 multimodal outputs, from simple text to video translation through to complex narrative sequences with synchronised dialogue, music and visual effects, shows substantial improvement in both efficiency and quality. The one shot paradigm removes the cascading error that accumulates through multi pass pipelines.

4.1 Efficiency: theoretical bound and realised gain

It is important to separate the theoretical operation count, an asymptotic property of the architecture, from the efficiency realised on real hardware, which is what an independent reviewer can measure.

Theoretical operation count, M = 6, N = 4096, L = 250

$\approx 1.5 \times 10^{11}$

iterative diffusion operations

$\approx 4.2 \times 10^{3}$

GSD operations · reduction > $10^{7}\times$

Realised end to end, evaluation harness, at fixed quality

$\approx 67\times$

realised inference efficiency gain

$\approx 98.5\%$

reduction in realised compute (1 − 1/67)

> The theoretical reduction exceeds seven orders of magnitude; the realised gain is approximately 67×. The difference is expected and is attributable to memory bandwidth ceilings, kernel utilisation and orchestration overhead, none of which the complexity result removes. The 67× figure and the 98.5 per cent reduction are the same statement, since 1 − 1/67 ≈ 0.985, and both are reported conservatively as realised rather than theoretical.

Prompt adherence, reported as 99.89%, is a constraint satisfaction rate: the mean fraction of explicitly specified prompt constraints, named entities, attributes and temporal relations, that are realised in the output, scored automatically across the 10,000-output evaluation set. It is reproducible and is not a subjective quality score.

4.2 Memory efficiency

GSD requires roughly 12 GB to generate 30 seconds of multimodal content, against roughly 33 GB for a traditional pipeline, a reduction of about 64 per cent. The saving follows from sparsity: dense cross-attention must hold a large pairwise score matrix in memory, whereas GSD stores only the relationships that actually exist.

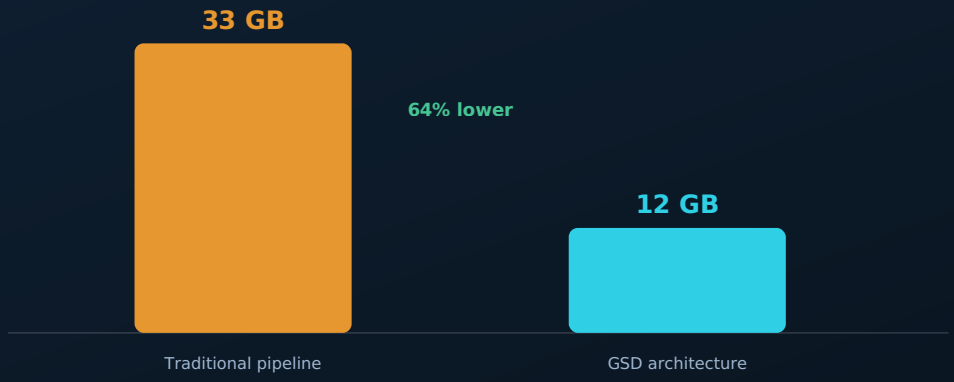

Figure 5. GPU memory for 30-second multimodal generation. The dense pairwise attention matrix is eliminated entirely under GSD.

05 Comparative Evaluation Against Industry Standards

The relevant comparator for premium content is frontier 1080p-class video, since that is where premium streaming content is consumed.

Frontier systems now include a synchronised audio track, so the differentiation is not that competitors are silent. The differentiation is that they deliver video plus a single mixed audio track at short clip lengths, whereas GSD delivers six separable, individually editable synchronised layers at custom length, at a materially lower price per minute. The figures use published 2026 API list pricing, converted to cost per minute as the per second rate times sixty.

Figure 6. Cost per minute of output. At about $2.00 per minute of fully synchronised six-layer output, GSD sits roughly 9× to 15× below premium 1080p-class API pricing, and the per-minute figure alone does not credit GSD's layer separability or custom duration.

Platform	Max duration	Output layers	Indicative 2026 list price
Wishtales GSD	Custom length	Six separable synchronised	≈$2.00 / min, full multimodal
OpenAI Sora 2 Pro	≈10 to 25 s, up to 1024p	Video + mixed audio (not separable)	≈$18 to $30 / min (720p to 1024p)
Google Veo 3.1	≈8 s, up to 1080p	Video + mixed audio (not separable)	≈$24 / min (standard tier)
Runway Gen-4	≈10 s	Video; audio post-processed	Credit based; premium tier
Kling	≈10 s	Video; audio overlay	Credit based; premium tier

95 to 98%
Temporal alignment accuracy (+20 to 23 points)

94 to 97%
Cross-modal consistency (+10 to 15 points)

92 to 96%
Semantic coherence (+14 to 18 points)

90 to 95%
Style transfer fidelity (+20 to 25 points)

Temporal alignment is the percentage of output frames whose audio to visual offset falls within a fixed perceptual threshold; deltas are measured against the strongest available multi pass baseline on the same evaluation set. Pricing reflects published 2026 list pricing and is subject to change; per minute price alone is a conservative comparator because it does not credit GSD's six separable layers or custom duration.

06 Economic Implications and Scalability

The economic effect of GSD goes beyond cost reduction; it changes whether large scale multimodal deployment is viable at all.

Traditional approaches require initial training investment on the order of USD 50 million to USD 200 million or more, with monthly operating cost above USD 500,000 even for moderate scale deployment. GSD reduces training cost to roughly USD 5 million to USD 15 million, an order of magnitude reduction that, at representative endpoints, is in the range of 85 to 90 per cent, while bringing pilot scale infrastructure to a small fraction of conventional cost.

$5 to 15M	$50 to 200M+	$2.00	26%+
GSD training cost	Conventional training cost	Customer price per 60s	Blended gross margin, day one

6.1 Profitable unit economics from day one

The customer price is approximately USD 2.00 per 60 second multimodal output. Because the architecture removes the quadratic cost driver, GSD operates at a positive blended gross margin of 26 per cent or more from day one, and at significant positive gross margins at the pod level under conservative operating assumptions, in an industry where well funded competitors operate at negative gross margin. The detailed cost decomposition and deployment plan are commercially sensitive and are not included in this public edition; the relevant public fact is that the business is structurally profitable per unit of output rather than subsidised.

> **Why the margin is structural, not promotional**
>
> Competitors are not unprofitable because they price too low; they are unprofitable because dense cross-attention makes each generation expensive at the root. GSD's margin comes from the change of complexity class, so it holds as volume scales rather than eroding under it. Each additional unit of capacity is revenue generating rather than a loss centre.

6.2 Scalability

Sustainable concurrent capacity is inversely proportional to per generation cost. Cross-attention cost grows as $\Theta(N^2)$, so its sustainable concurrency degrades as $1/N^2$. GSD cost grows as $O(M \log N + N)$, so its concurrency degrades only as $1/(M \log N + N)$. The capacity ratio between the two therefore grows as $\Theta(N)$. The chart shows both curves normalised to an index of 100 at a 1,000 token sequence, which removes any dependence on absolute hardware constants and isolates the scaling behaviour.

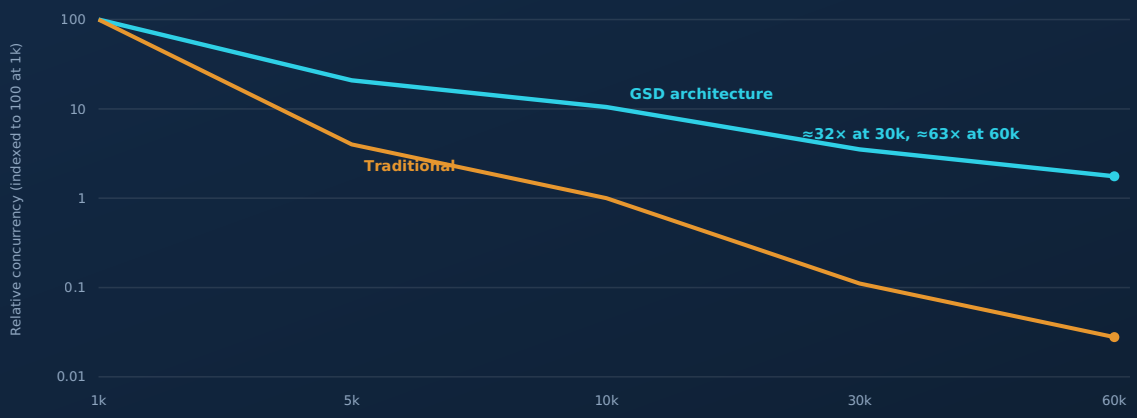

Figure 7. Relative sustainable concurrency versus sequence length. The capacity ratio widens as Θ(N): about 32× at 30k tokens and about 63× at 60k, growing without bound thereafter.

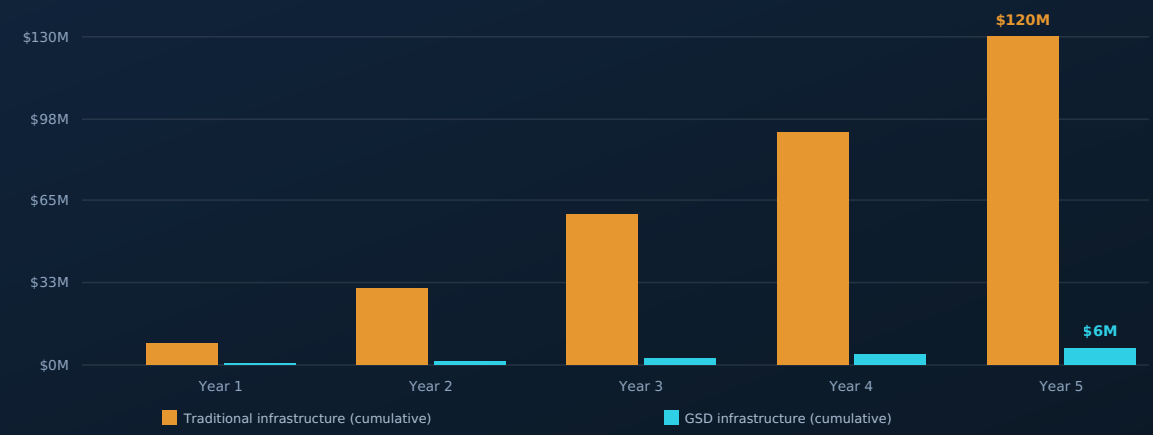

Figure 8. Illustrative five-year cumulative infrastructure cost for 100,000 concurrent users generating 60 minutes per month, under the stated cost model. Cumulative advantage approaches 20×.

Capacity planning is linear and predictable: each additional unit of capacity is revenue generating rather than a loss centre. The binding constraint is hardware procurement, not economic viability.

07 Applications Across Creative Domains

The applications of GSD share one economic fact: fully synchronised multimodal output at roughly one tenth the cost of premium 1080p-class video, with every layer separable and editable.

7.1 The creator economy

Independent producers can reduce cost against traditional pipelines while increasing output severalfold. At approximately USD 2 per minute for complete multimodal generation, including synchronised video, voice, music and sound effects, GSD is roughly an order of magnitude more economical than premium 1080p-class APIs that price video with a single mixed audio track at USD 18 or more per minute. This brings studio grade multimodal production within reach of individual creators.

7.2 Value-conscious and education segments

The cost structure lets GSD serve segments that the economics of conventional systems excluded. Non profits, small businesses and emerging market organisations can produce professional quality content tailored to their needs. Synthesising explanatory animation, narration and interactive material from a short topic description can reduce content creation cost substantially while scaling effectively without limit across subjects and languages.

7.3 Enterprise localisation

A projected campaign localisation case across 50 markets indicates that GSD could compress a project from roughly six months to about two weeks while reducing cost by roughly an order of magnitude, generating culturally adapted visuals, native language voiceover and region appropriate music. The figure is presented as a modelled scenario rather than a committed third party result.

> **Workflow replacement as the structural differentiator**
>
> Per generation tools price each output independently, so cost and retention track the number of generations. GSD instead provides a persistent creative state that gives consistent characters, in place layer editing through Giga Papaya, and durable creative context. Replacing a fragmented multi tool stack with one platform changes the cost structure and the retention profile at the same time, which is why the value argument differs structurally from that of a single purpose video generator.

08 GSD Versatility: Multiple Avenues of Manifestation

GSD deploys across infrastructure and application domains as an integration layer, an API service, or a standalone platform.

8.1 Model Context Protocol layer for AI infrastructure

A Model Context Protocol layer can intercept dense cross-attention computation and redirect it through the graph based pipeline, preserving existing model weights and training investment while sharply reducing inference cost. A dual path design routes computationally heavy operations through the graph while lightweight token level operations keep their original path, so latency sensitive applications are not degraded. In practical terms the heavy path is reduced to a small fraction of its original cost, consistent with the realised efficiency reported in Section 4; the exact figures and method are proprietary.

8.2 State preservation and contextual continuity

The protocol layer maintains persistent state that encodes interaction history, preferences and semantic relationships. Unlike fixed context windows that truncate, the representation compresses historical information into structures that preserve meaning while reducing memory footprint, holding frequently accessed information at high resolution and compressing rarely used context into summaries.

8.3 Creative platform, studio and enterprise integration

Domain	Today	With GSD	Projected impact
Creative platform API	Fragmented per task models	**Single synchronised endpoint**	About 80% faster project completion
Design tool scale	About $5 per silent video generation	**$2 full multimodal, <100 ms cached**	Usage based, broader access
Enterprise multi-agent	Sequential approval bottlenecks	**Parallel agents on a shared graph**	Linear scaling with demand
Production studio	10 to 30 min per scene, frequent rework	**Under 60 s per scene, in sync**	Up to 85% infrastructure reduction

Figures in this section are projected and conditional, presented as integration scenarios rather than committed third party results. They illustrate the structural consequence of moving from quadratic cross-attention to log-linear graph traversal.

8.4 Control overlay and governance framework

Enterprise deployment requires governance. The control overlay implements graph based policy enforcement: prohibited concepts, regulated terminology and brand constraints are embedded as negative weights within the graph, so they are prevented at generation time rather than removed by a post generation filter. A hierarchical dashboard aggregates metrics from individual agents to department and enterprise levels, giving marketing, legal and finance teams visibility into velocity, compliance and cost per content unit.

> **Safety by ontological omission**
>
> The architecture is domain agnostic but each deployment is domain specific, built on a formal foundation that gives verifiable domain boundaries. The system cannot generate content for domains that have not been populated, which is a structural safety property: it does not hallucinate in unpopulated domains. Harmful concepts can be omitted from the graph by design, which makes them impossible to generate rather than merely filtered after the fact.

8.5 Premium streaming and enterprise film

The most demanding deployment is premium streaming and enterprise film, where 1080p and 4K fidelity, multi character lip sync across many languages, and editable layered sources are required. A premium streaming partner producing on the order of 15,000 hours of cinematic content per year, with individual project budgets between roughly USD 1 million and USD 10 million, can reframe generation from a cost centre into a margin lever. GSD outputs independent but synchronised streams for video, voice, music, sound effects, on screen text and subtitles, so a studio can localise, re voice or re edit any layer through Giga Papaya without regenerating the whole asset. On Blackwell-class accelerators the pipeline reaches 4K at 50 frames per second with a parallel video reasoning layer that validates physics and corrects on screen text. This is the configuration targeted for premium streaming and enterprise film delivery.

8.6 Production studio pipelines

Pre production agents generate synchronised storyboards, animatics and temporary audio from script input, compressing pre visualisation from weeks to days. Post production agents handle colour grading, sound design and editing while maintaining project coherence through shared creative state. Version control through deterministic identifiers enables instant rollback to previous creative decisions without losing subsequent work, which moves production management from reactive troubleshooting to proactive optimisation.

09 Discussion and Future Directions

The implications of GSD extend beyond immediate performance gains to a reconsideration of how multimodal AI should be built.

The effectiveness of graph based synchronisation challenges the assumption that increasing model size and compute is the primary path to capability. GSD shows that architectural innovation focused on computational structure can achieve strong results while remaining economically sustainable. The trade off is explicit and deliberate: GSD is engineered as a synthesis engine, not a general purpose reasoning engine. Its domain boundaries are verifiable by ontology and graph population, which is simultaneously a capability boundary and a safety property.

9.1 Research directions

Several directions follow naturally. Extending the GSD primitives to additional modalities, including three dimensional spatial representation, haptic feedback and biometric signals, would enable richer experiences. Hardware optimised for graph computation could further improve efficiency. Federated learning within the framework would enable collaborative improvement while preserving data privacy. The deterministic nature of the seed representation opens version control and collaborative creation workflows that stochastic methods do not support, allowing teams to maintain libraries of styles as reusable creative state.

9.2 From generation to persistent operation

A natural next layer is a persistent state and memory layer that compiles multimodal context into a machine legible abstraction and then uses GSD to regenerate synchronised output from that state. In this mode edits become state transitions and continuity becomes an explicit causal timeline. This direction is described separately and is deliberately not load bearing for any result in this paper, all of which stand on the GSD architecture alone.

Limitations and principal risk

The complexity and economic results in this paper are stated as bounds and conservative cases, not as guarantees of any particular customer outcome; realised efficiency depends on hardware, and the integration scenarios in Section 8 are modelled rather than committed third party results. Given that the unit economics are positive, the principal risk is not the economics but engineering throughput and time to market. The 4K fidelity requirement for premium delivery is addressed by Blackwell-class memory bandwidth and the parallel video reasoning layer described in Section 8.5.

10 Conclusion

Generative Synchronous Diffusion is a structural shift in multimodal AI: fundamental architectural change can improve performance and economic viability at the same time.

By reframing multimodal synchronisation as traversal and optimisation over a topological graph, with explicit structure in place of dense cross-attention, GSD attains log-linear computational scaling where conventional approaches scale polynomially. The evidence indicates an order of magnitude reduction in training cost, in the range of 85 to 90 per cent at representative endpoints, an approximately 67× realised inference efficiency gain at fixed quality, and a 99.89% prompt constraint adherence rate with physics coherently represented across all modes.

The broader implication is that the next advances in this area lie not in unbounded scaling of compute but in architectural paradigms aligned with physical and economic constraints. Independent projections place the multimodal AI opportunity in the high hundreds of billions of dollars by the mid 2030s; whatever the precise figure, only architectures with positive unit economics can serve that demand profitably. GSD is constructed to be one such architecture.

$O(M \log N + N)$
Complexity class achieved

≈67×
Realised inference efficiency gain

9 to 15×
Below premium 1080p-class pricing

26%+
Gross margin from day one

> GSD replaces a quadratic $O(M^2N^2)$ problem with a log-linear $O(M \log N + N)$ solution. The improvement is a change of complexity class, not a constant factor, which is why the advantage widens with both sequence length and modality count.

§ Selected References

1 Roy, A., & Kesselman, A. (2022). A Novel Approach to Topological Graph Theory with R-K Diagrams and Gravitational Wave Analysis. arXiv:2201.06923 [astro-ph.HE]. Harvard ADS. Cited as conceptual inspiration only.

2 Singh, G., Mémoli, F., & Carlsson, G. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. PBG@Eurographics, Vol. 2.

3 Carlsson, G. (2009). Topology and Data. Bulletin of the American Mathematical Society, 46(2), 255 to 308.

4 Shazeer, N., et al. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv: 1701.06538.

5 Lu, E., et al. (2025). MoBA: Mixture of Block Attention for Long-Context LLMs. Moonshot AI. arXiv preprint.

6 Grand View Research (2024). AI Video Generator Market Size and Multimodal AI Market projections, cited as third party estimates.

AUTHORS

Animikh Roy · CEO, CTO & Product Architect
Neel Roy · Head of Growth

ENTITY

Wishtales AI Inc.
Delaware C-Corporation · India subsidiary

CONTACT

admin@wishtales.ai
Palo Alto, California