GENERATIVE SYNCHRONOUS DIFFUSION · GSD

The Next Breakthrough in Multimodal AI

For Creators, Agents and Enterprises

A fourth architectural pillar for generative AI, after attention, diffusion transformers and state space models: a topological graph foundational model that collapses the multimodal scaling wall from $O(M^2N^2)$ to $O(M \log N + N)$ and generates six synchronised, independently editable layers in one shot, now at 120 seconds and beyond in 4K at 50 frames per second, rendered on Blackwell B300 in about twice the running time, with positive unit economics from day one.

WHAT THIS PAPER ESTABLISHES

120 s+
One shot output, no camera cuts, six layers in sync

4K · 50 fps
Native output with coherent physics and visual effects

98.99%
Prompt to physics adherence across the timeline

≈67×
Realised inference efficiency gain at fixed quality

$2.00
Per 60 s, full multimodal, edits included

50%
Gross margin on pure GSD output, day one

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

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

Public Technical Overview · Version 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), a topological graph foundational model that reframes multimodal synthesis as retrieval and combinatorial optimisation over an explicit world graph followed by a single linear render pass, replacing the polynomial scaling of cross attention and iterative denoising with log linear scaling. It is positioned as a fourth architectural pillar of generative AI, distinct from attention, from the diffusion transformer beneath every frontier video model, and from the state space model, whose linear time recurrence it shares while replacing the learned latent state with a quasi deterministic topological blueprint. Across six synchronised modalities the dominant per generation cost falls from $O(M^2N^2)$, quadratic in both the modality count M and the sequence length N, to $O(M \log N + N)$. This Version 2026 edition records what that change of complexity class delivers at production scale on Blackwell B300 accelerators: a single one shot pass produces 120 seconds and beyond of 4K video at 50 frames per second in about 240 seconds, with voice, lip sync, sound effects, background music and on screen text generated at once and in sync, without a camera cut, and with physics and visual effects coherently represented across every layer. On the evaluation harness this corresponds to an approximately 67× realised inference efficiency gain at fixed quality, 99.89% prompt constraint adherence and 98.99% prompt to physics adherence. Giga Papaya, the layer selective editing engine on the same architecture, edits any layer at one second granularity without regenerating the asset. The result is the fastest, least expensive and most physically faithful long form generation the field has published, at USD 2.00 per minute with a 50 per cent gross margin inside the price, against USD 24 to USD 42 per minute for the like for like premium tiers, 1080p and 4K with a mixed audio track, in clips of eight to thirty seconds. Because the same core is exposed by API, GSD and Giga Papaya can serve as the generative back end beneath the creator platforms that today resell third party clip models. 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), in which topological signatures identified gravitational wave merger events without computing posteriors. That research is an inspiration, not a foundation: only the R-K toolkit, representing complex relationships as R-K diagrams and optimising combinatorially over them, served as a starting point. GSD is an independent architecture; its design, optimiser and signature construction are proprietary and not described here.

Contents

1 Introduction: The Economics of Multimodal Intelligence
2 Theoretical Foundation and the Scaling Wall
3 Architecture: A Paradigm Shift in How Synthesis Works
4 Performance: Efficiency, Speed and Quality
5 Comparative Evaluation Against Industry Standards
6 Economic Implications and Scalability
7 What Creators Need: The Four Cs
8 Applications Across Creative Domains
9 Intellectual Property: Six Patent Families
10 GSD Versatility: Multiple Avenues of Manifestation
11 Discussion and Future Directions
12 Conclusion

About this edition. *This public technical overview, Version 2026, supersedes the earlier public overview and records the advances since it: one shot generation at 120 seconds and beyond in 4K at 50 frames per second on Blackwell B300, the Giga Papaya per second editing engine, the fourth pillar positioning, the benchmark position against the commercial frontier, the six patent families on file, and the updated unit economics. Wishtales figures are measured and conservative; competitor figures come from public documentation and list pricing and are subject to change.*

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, about 32 per cent compound annual growth over the 2025 to 2033 window (Grand View Research). These 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 as $O(M^2N^2)$, quadratic in both. For a system that must synchronise six modalities simultaneously this 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.

$O(M^2N^2)$	$O(M \log N + N)$	$M = 2 \rightarrow 6$	50%
The scaling wall competitors face	What GSD achieves	Synchronised modalities, in production	Blended gross margin on pure GSD output, 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.

What Changed Since the Previous Edition

Dimension	Previous public overview	Version 2026
Single pass output	Custom length; 4K 50 fps targeted	**120 s+ at 4K 50 fps, no cuts, six layers in sync, about 2× real time**
Adherence	99.89% prompt constraint adherence	**99.89% constraint plus 98.99% physics adherence**
Editing	Layer swap without full re render	**Any layer, any second, propagated and reflowed; edits included**
Unit economics	$2.00 per minute; 26% or more gross margin	**$2.00 per minute; 50% pure, 35 to 40% with unlimited edits**
Intellectual property	Proprietary, undisclosed	**Six patent families filed (Section 9)**

1.1 The Field Has Reset

The period since the first edition of this overview has confirmed the thesis in the least comfortable way possible: the cost structure of dense cross attention is now visible in the product decisions of the best funded companies in the field.

- **OpenAI has discontinued Sora.** The Sora web and app experiences have been closed and the Sora API is being retired, with the company citing compute cost and a reallocation of the research team toward world simulation for robotics. Disney's USD 1 billion agreement built around Sora was dissolved at the same time. The most heavily promoted video model in the industry was withdrawn because its unit economics could not be made to work.
- **Google's Veo line has not advanced architecturally.** Veo remains on the 3.1 generation with no successor announced. Its public additions have been product features such as scene extension and 4K upscaling rather than a new architecture, and its native ceiling remains an eight second clip.
- **Runway, Pika and Luma have published no fundamental architectural research.** Their own models remain short clip diffusion transformers; Luma has moved to an agentic interface over existing models, and Runway's flagship still renders silent 720p clips of five or ten seconds through its API.
- **The United States creator platforms have become distribution layers for Chinese models.** Runway, ElevenLabs, Higgsfield, OpenArt, Krea and Freepik now route video generation to ByteDance's Seedance and Kuaishou's Kling; Runway bundles Seedance 2.5 in its top plan and resells it by the second through its API, and ElevenLabs offers Seedance and Kling alongside Veo. The model that holds the thirty second frontier, Seedance 2.5, is Chinese. The interface layers are American; the generative core increasingly is not.

Every model in that list, American or Chinese, is built on the same quadratic primitive. The frontier has moved from eight to thirty seconds of native clip length, priced at roughly USD 24 to USD 42 per minute at the like for like 1080p and 4K tiers with a mixed audio track, and none of it changes the complexity class. GSD is a different primitive, and Section 1.2 places it as a fourth architectural pillar beside the three the field already recognises. To the authors' knowledge it is the only production architecture that generates a two minute, six layer, 4K asset in one pass, and the only one that does so at a profit: a domestically owned, patented generative core that the platforms above can call by API in place of the clip models they resell today, rather than another interface over someone else's diffusion transformer.

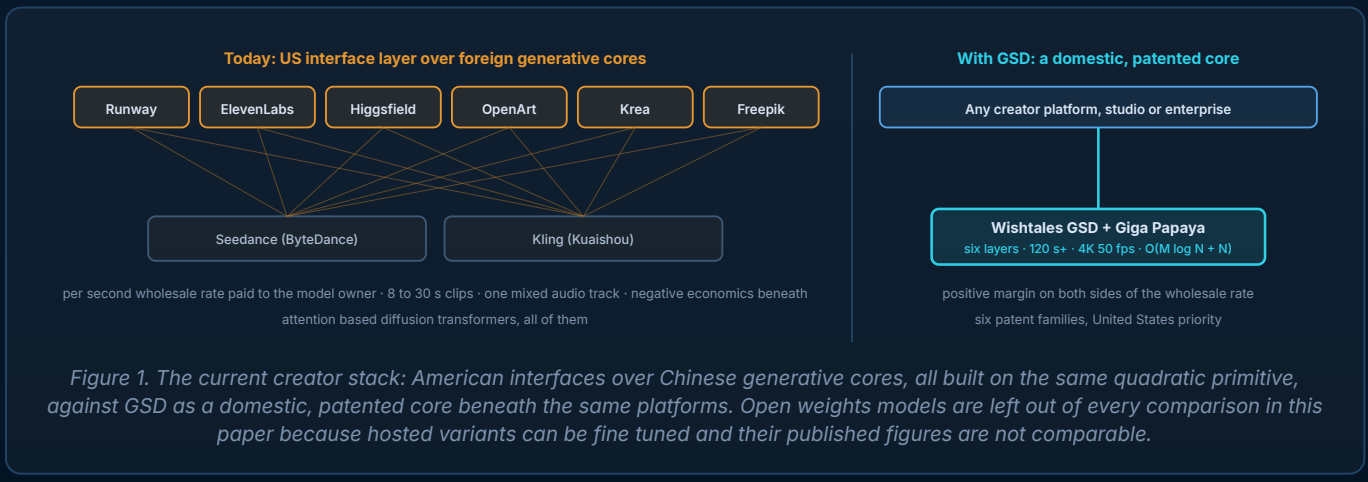


Figure 1. The current creator stack: American interfaces over Chinese generative cores, all built on the same quadratic primitive, against GSD as a domestic, patented core beneath the same platforms. Open weights models are left out of every comparison in this paper because hosted variants can be fine tuned and their published figures are not comparable.

What Creators Actually Ask For

Creators, agencies and studios ask four things of a video system: it must be **convenient**, its output must be **consistent**, it must give **creative control**, and it must be affordable in **cost**. Existing systems trade these against each other; a longer clip costs more, a consistent character needs reference wrangling, and an edit means a full regeneration. Section 7 sets out how GSD answers all four from a single architectural decision, and Section 9 sets out the six patent families that protect it.

1.2 The Fourth Pillar: A Topological Graph Foundational Model

Generative AI has so far rested on three architectural pillars. **Attention**, the transformer primitive, computes pairwise interactions between every pair of tokens; it is the engine of language models and the conditioning path of every video model, and it is quadratic in sequence length. The **diffusion transformer** puts a transformer backbone inside an iterative denoising loop; it is the architecture of every frontier commercial video system, Seedance, Veo, Kling, Runway and the discontinued Sora alike, and it pays the attention cost again on every one of its denoising steps, which is why its native horizon has stalled at eight to thirty seconds and its output is one flat file. The **state space model** replaces attention with a linear recurrence over a learned latent state; it is linear in sequence length and strongest in speech and audio, but its state is an opaque vector and it generates one stream at a time.

GSD is a fourth pillar. It is an independent approach rather than a variant of the third, although it sits closest to it: the render stage is a single linear pass in which every time step combines the tensor summary of the previous step with the blueprint for the current one, much as a multi head state space model projects one state into several output streams. What is new is the state itself. In GSD it is not a learned latent but an explicit, quasi deterministic topological blueprint, retrieved from a world graph and resolved by combinatorial optimisation before a single frame is rendered. There is no denoising loop and no UNet, and the optimisation stage is combinatorial rather than gradient based: the pipeline already knows what it is going to draw, so compute goes into filling the shape rather than discovering it. That is the property that yields six synchronised layers instead of one stream, a horizon bounded by memory rather than by cost, and an output that can be edited to the second.

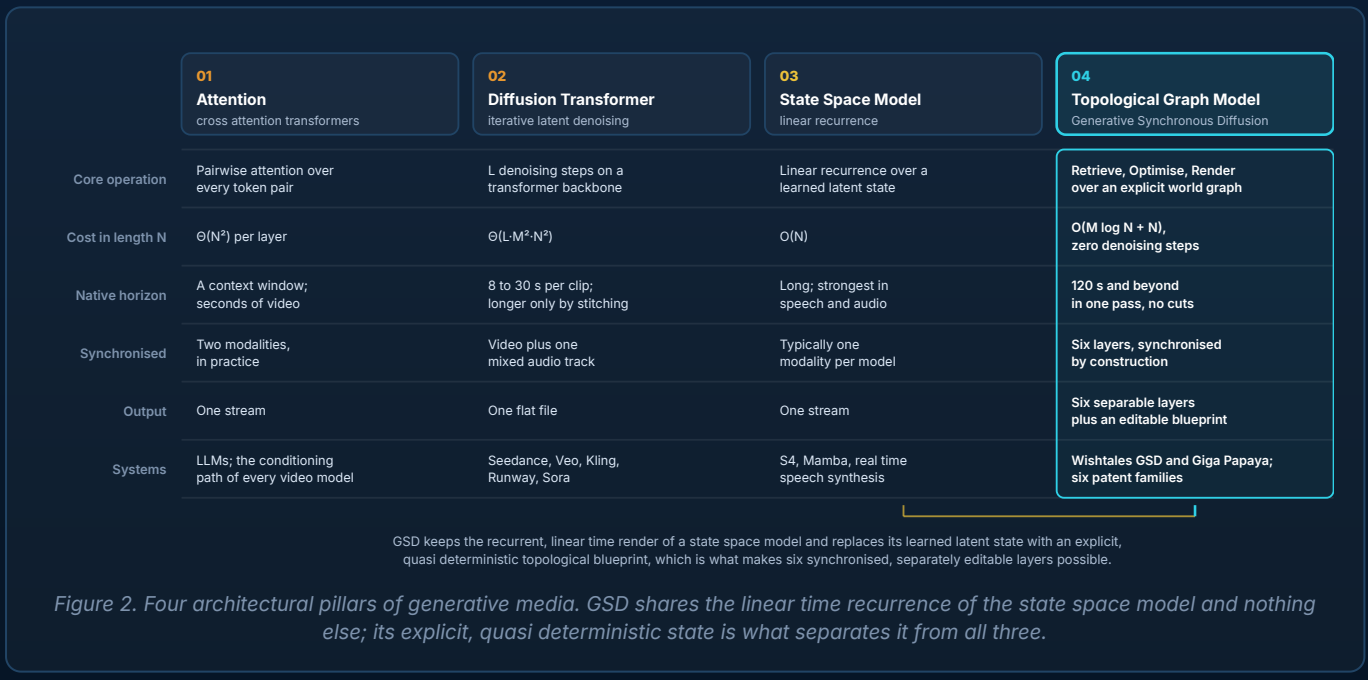


Figure 2. Four architectural pillars of generative media. GSD shares the linear time recurrence of the state space model and nothing else; its explicit, quasi deterministic state is what separates it from all three.

Question	Attention	Diffusion transformer	State space model	Topological graph model (GSD)
How is length paid for	Quadratically, per layer	Quadratically, L times over	Linearly	**Linearly, after a one time log cost lookup**
Where is the world	In the weights	In the weights	In a latent vector	**In an explicit graph with topological invariants**
How is sync achieved	Cross attention, pairwise	Joint denoising, one mixed track	Not addressed across modalities	**By construction, one blueprint, six heads**
What is the horizon	The context window	8 to 30 s per clip	Long, single stream	**120 s and beyond in one pass, demonstrated to 154 s**
What comes out	One stream	One flat file	One stream	**Six separable layers and the blueprint that made them**

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, lip sync, sound effects, music and on screen text), each rendered as a separable layer from one shared state.
N	Sequence length, that is, the number of text tokens or image and video patches, typically between 77 and a few thousand for a clip and tens of thousands for a 120 second 4K timeline.
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. It is also why the field's native clip length has crept from eight to thirty seconds over a year rather than leaping to minutes: every extra second is paid for quadratically.

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 quasi 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	**Quasi 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.

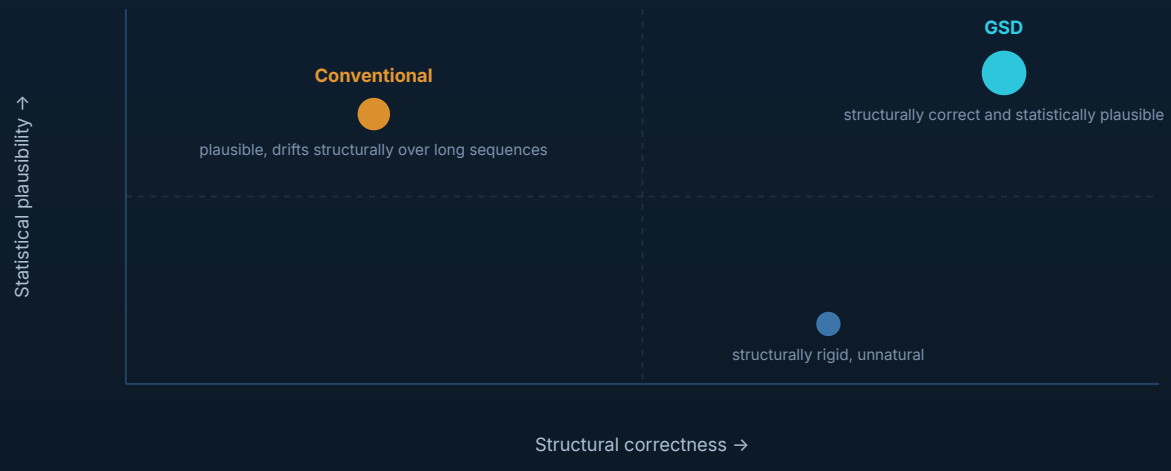


Figure 3. Quality is judged on two independent axes. GSD targets the upper right: structurally correct and statistically plausible at the same time, which is where consistency over 120 second sequences comes from.

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.

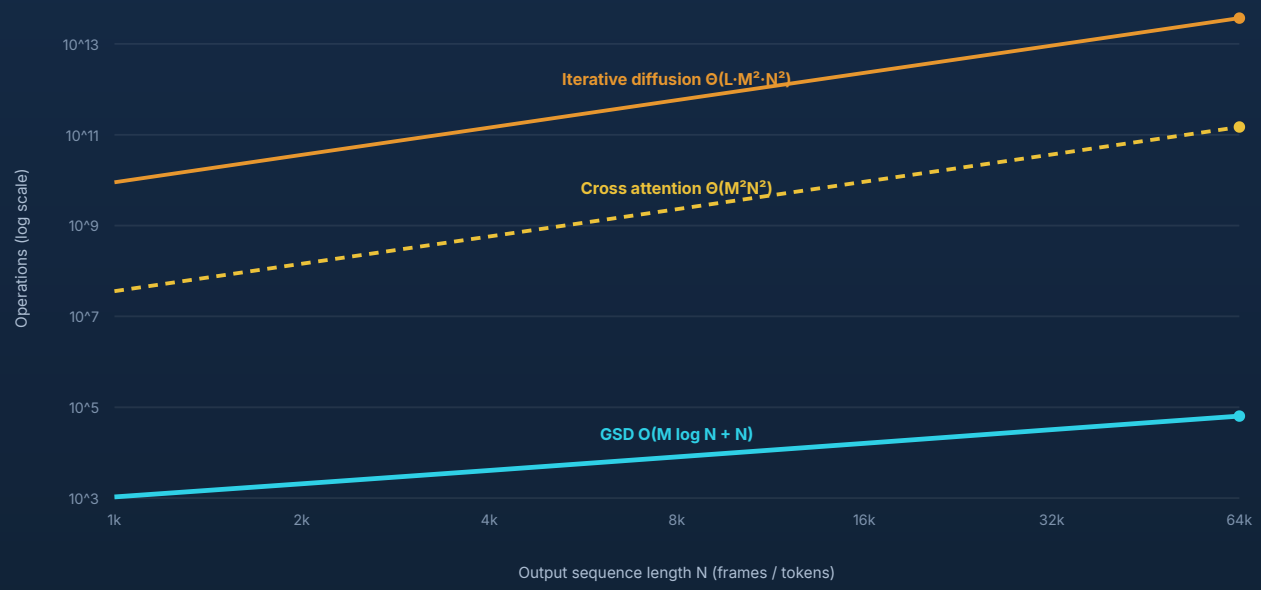


Figure 4. Theoretical operation count versus output sequence length. Cross attention and iterative diffusion scale polynomially; GSD scales log linearly.

N (sequence)	Cross attention $\Theta(M^2N^2)$	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. A 120 second timeline at 4K sits at the right hand end of this chart, which is why the field's clip ceiling stops at thirty seconds and GSD's does not. Realised hardware efficiency is lower than the theoretical bound and is reported conservatively in Section 4.

Where a 120 Second, 4K Timeline Sits on This Axis

At 50 frames per second a 120 second output comprises 6,000 frames, and at 4K each frame carries an order of magnitude more patches than a 720p frame. In sequence terms that places a single GSD output at the far right of Figure 4 and beyond, where the quadratic curves have already left the range that any price could recover. It is the same reason the field's native clip ceiling has settled at eight to thirty seconds, and the reason GSD's has not: a change of complexity class is the only thing that moves that ceiling by an order of magnitude rather than by a few seconds.

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 render stage is a set of modality projection heads reading one shared, time indexed state, much as a multi head state space model projects one state into several output streams; GSD is not several diffusion streams sharing an attention window. 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 for the whole timeline. 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 through six projection heads. 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.

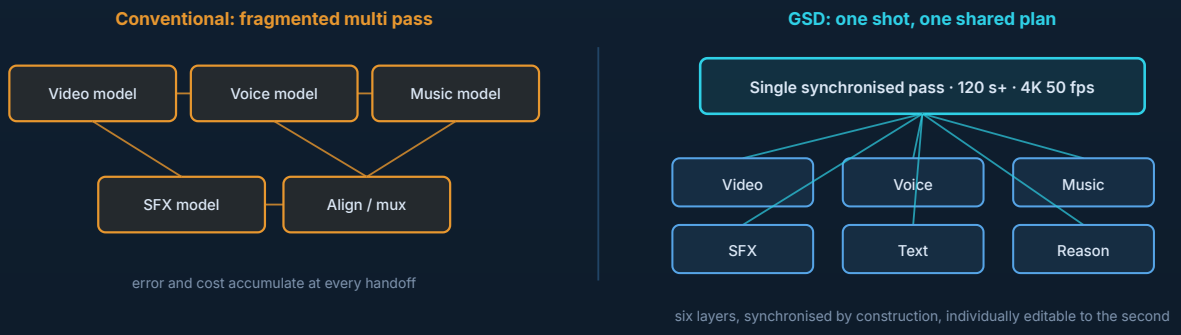


Figure 5. The architectural inversion. Conventional pipelines align after generation; GSD synchronises before it, so the six output layers cannot drift relative to each other, at any length.

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, at fifteen seconds or at two minutes. 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. Frontier systems now 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.

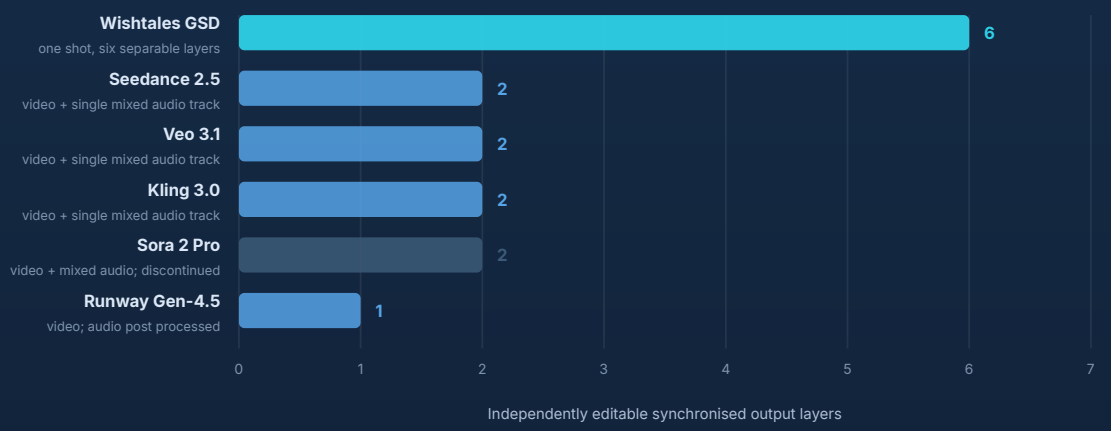


Figure 6. Independently editable synchronised output layers, an approximate characterisation based on publicly described capabilities. Competitor audio is a single mixed track rather than separable voice, music and effects stems.

3.4 Long Form in One Shot: 120 Seconds and Beyond

Since the first edition, the ceiling of a single GSD pass has moved from tens of seconds to 120 seconds and beyond, at 4K and 50 frames per second, with no camera cut required to reach that length. The advance is a consequence of the architecture rather than a new model: because the plan for the whole timeline is resolved once in the optimisation stage and the render is linear in N, extending the output extends the render pass rather than compounding an attention cost. Every second of the output is derived from the same shared state, so a character who appears at second 3 is the same character at second 118, with the same voice, and the physics that governed the first shot governs the last. The field's native ceilings remain 8 to 30 seconds per clip, with Seedance 2.5 holding the thirty second frontier; longer sequences are assembled from separate generations, each of which reintroduces drift at the boundary and a fresh charge per second. The longest single GSD pass demonstrated to date is a 154 second commercial with no cut and no stitch.

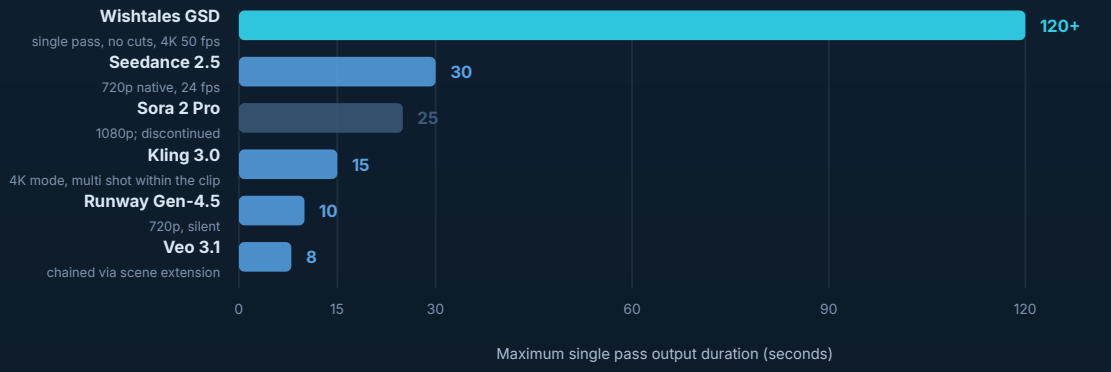


Figure 7. Maximum single pass output duration by system, from publicly described capabilities. GSD is shown at 120 seconds; longer single pass outputs are a function of accelerator memory headroom rather than of the architecture.

3.5 Editing Without Re rendering: Giga Papaya

Because the six layers are separable and every second of every layer is addressed on one shared timeline, Giga Papaya lets a user change one layer, for example re voice a character, swap the music, alter a costume or extend a shot, without regenerating the whole asset. Conventional systems must re render the entire clip for any change.

Giga Papaya is a localised, layer selective editing engine for synchronised multimodal composition. It resolves an edit to a bounded slice of a temporally indexed structural blueprint through temporal topological localisation, quantised to every second of every layer, and re renders only the implicated modality projection head for only the implicated seconds in a quasi deterministic manner, leaving every unedited layer untouched. A visual reasoning layer detects the edited entity where it recurs in later scenes and propagates the change on user confirmation. Duration changing edits automatically reflow the timing of the remaining layers, so a line of dialogue that runs a second longer moves the music and the effects with it. The cost of an edit therefore scales with the extent of the edit, not with the length of the asset.

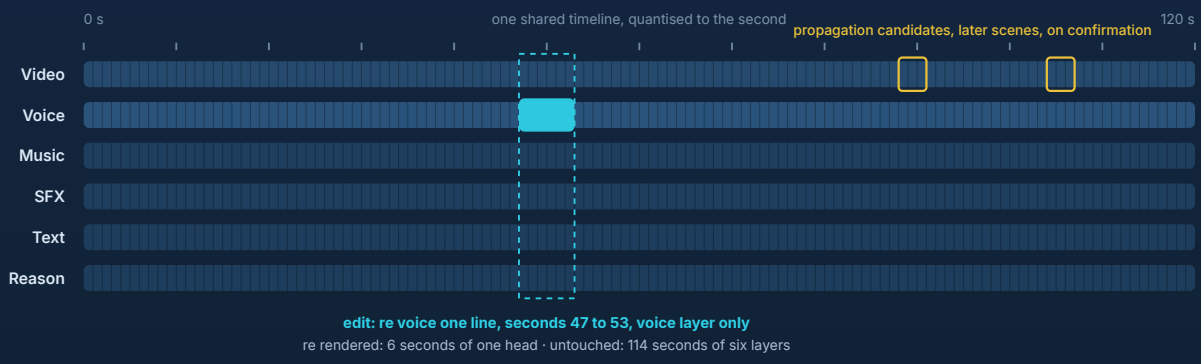


Figure 8. A Giga Papaya edit on a 120 second, six layer timeline. One line is re voiced between seconds 47 and 53; only the voice head is re rendered for those seconds, every other second of every other layer is untouched, and later occurrences of the same entity are offered for propagation.

1 s

Edit quantisation, per layer. In consumer terms Giga Papaya is to a synchronised video timeline what conversational image editors are to a single still: describe the change, and only that changes.

65 to 90%

Compute saved on a typical edit against full regeneration, which is why an effectively unlimited number of edits can sit inside the USD 2.00 per minute price rather than being metered.

A Structural Advantage of the One Shot Architecture

Layer selective, second quantised editing is only possible because the layers were generated as separable projections of one shared plan in the first place. A system that emits a single mixed video file has nothing to localise an edit to. The mechanism is proprietary and is protected as its own patent family (Section 9).

04 Performance: Efficiency, Speed 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 efficiency, speed 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

≈1.5×10¹¹

iterative diffusion operations

≈4.2×10³

GSD operations · reduction > 10^7×

REALISED END TO END, EVALUATION HARNESS, AT FIXED QUALITY

≈67×

realised inference efficiency gain

≈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.

4.2 Adherence: Prompt Constraints and Physics

99.89%

Prompt constraint adherence. 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.

98.99%

Prompt to physics adherence. The fraction of physically specified or physically implied constraints in a prompt, such as mass and momentum, gravity, collisions, occlusion, fluid and cloth behaviour, lighting and shadow, that are realised without a violation at any frame across the full timeline, scored by the parallel video reasoning layer on the same evaluation set.

The physics metric is the stricter of the two because a single violated frame fails the constraint for the whole asset, and it is scored over 120 second outputs rather than eight second clips. Physics and visual effects are represented coherently across all six layers: a sound effect fires on the frame of the impact it describes, and the voice and music layers hold their timing when the visual physics dictates a longer or shorter action. Explicit structure is what makes this possible; the world graph carries the relationships that a pixel model has to rediscover in every frame.

4.3 Speed: Faster Than the Field, at Six Layers and 4K

Because the render stage is a single linear pass over a plan that has already been resolved, wall clock generation time grows in proportion to the output rather than compounding with it. On NVIDIA Blackwell B300 accelerators at FP8 precision, the production configuration, a fifteen second, six layer, 4K asset at 50 frames per second is generated in about thirty seconds and a 120 second asset in about 240 seconds; the longest single pass demonstrated to date, a 154 second commercial, rendered in about 300 seconds. The ratio of output seconds to wall clock seconds holds at about 0.5 across the whole range, because the plan is resolved once and the render cost per second does not change with the length of the timeline.

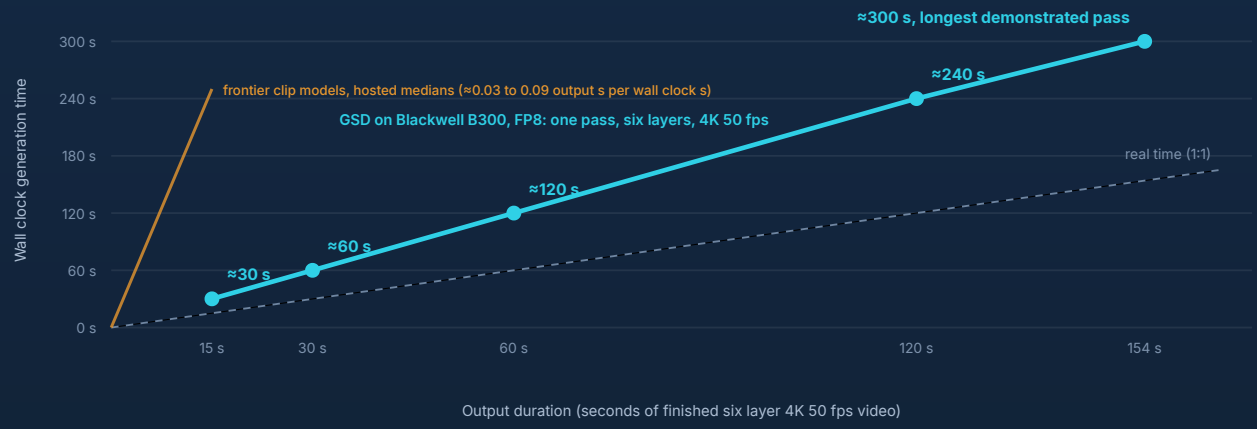


Figure 9. GSD wall clock generation time against output duration on Blackwell B300 at FP8. The dashed line is real time; the orange segment is the published hosted median for frontier clip models, which reaches fifteen seconds of output only after roughly three to eight minutes. Every GSD point is a complete six layer 4K asset at 50 frames per second.

Output duration	GSD wall clock	Output s per wall clock s	Advantage over published frontier clip medians
15 s	**≈30 s**	**0.50**	≈5 to 17×
60 s	**≈120 s**	**0.50**	≈5 to 17×
120 s	**≈240 s**	**0.50**	≈5 to 17×
154 s (longest demonstrated pass)	**≈300 s**	**0.51**	≈5 to 17×

Published thirty day median wall clock times for one eight second, 720p to 1080p clip on the current frontier run from roughly ninety seconds to four minutes on hosted platforms (Kling about ninety seconds, Veo 3.1 under two minutes, Seedance three and a half to four minutes), a ratio of about 0.03 to 0.09 output seconds per wall clock second for one video layer with a mixed track, and proportionally longer at the thirty second maximum; aggregator reports placed the discontinued Sora 2 Pro at thirty to ninety seconds for a five to ten second clip. GSD's ratio is five to seventeen times higher than the active frontier while producing six separable layers at 4K and 50 frames per second, and a 120 second asset arrives as one pass rather than as four to fifteen clips and an assembly step. Competitor timings vary by queue depth and are indicative only; on the previous accelerator generation the same GSD pass takes several minutes, so the Blackwell figures are the production reference.

Why Generation Time Is Linear in the Output

In a cross attention pipeline every additional second of output enlarges the pairwise score matrix that every denoising step must compute, so generation time grows faster than the output. In GSD the plan is resolved once and the render is one linear pass through six projection heads, so a longer output is simply a longer pass and the advantage widens with duration. Fastest, least expensive and most physically faithful are one fact seen three ways: the pipeline never rediscovers the scene at every step.

4.4 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. The same property is what lets a 120 second timeline fit within the memory of a single Blackwell class accelerator.

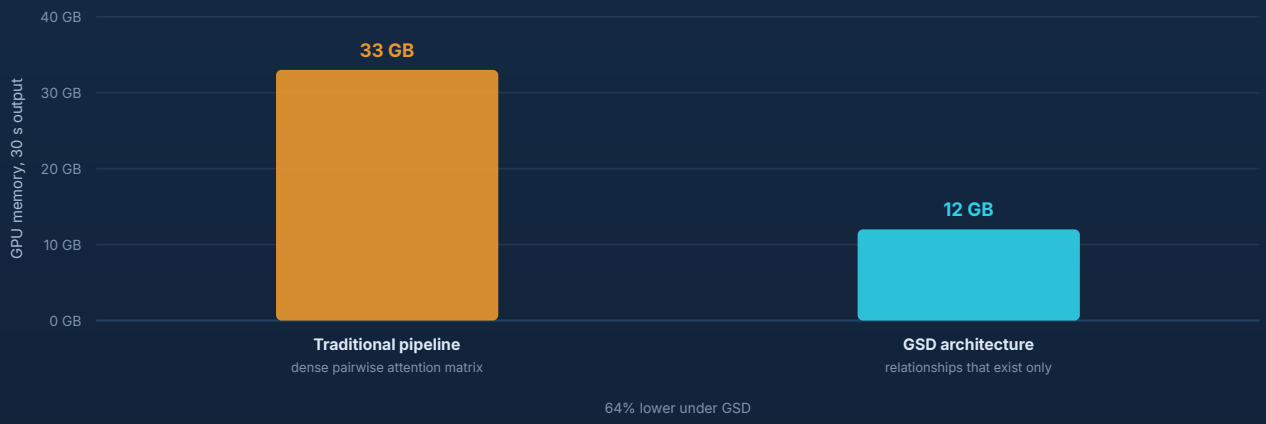


Figure 10. GPU memory for 30 second multimodal generation. The dense pairwise attention matrix is eliminated entirely under GSD.

4.5 Quality Deltas Against the Strongest Multi Pass Baseline

Beyond efficiency, the one shot paradigm improves the output itself. The four measures below are scored on the same 10,000 output set; each delta is measured against the strongest available multi pass baseline assembled from frontier clip generation, separate voice synthesis and post hoc alignment.

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)

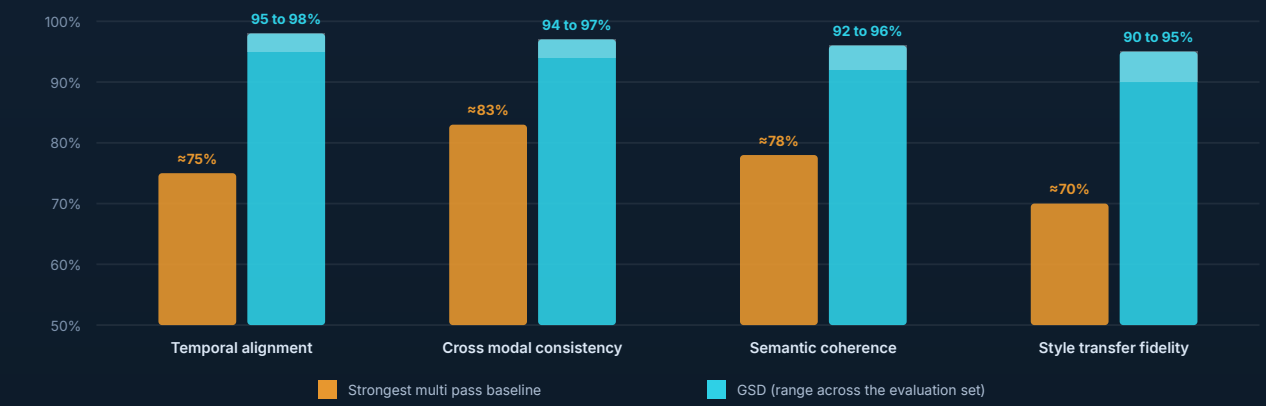


Figure 11. Quality deltas against the strongest available multi pass baseline on the same 10,000 output set. GSD bars show the observed range; the baseline is the midpoint implied by the stated deltas.

Temporal alignment is the percentage of output frames whose audio to visual offset falls within a fixed perceptual threshold; cross modal consistency measures agreement between what is seen, heard and written on screen; semantic coherence measures adherence to the narrative and scene logic across the whole timeline; style transfer fidelity measures retention of a specified visual and sonic style from the first second to the last. Deltas are measured against the strongest available multi pass baseline on the same evaluation set.

05 Comparative Evaluation Against Industry Standards

The relevant comparator for premium content is frontier 1080p to 4K class video with native audio, since that is where premium streaming content is consumed.

Frontier systems now include a synchronised audio track, and the best of them reach thirty seconds per clip, so the differentiation is neither silence nor brevity alone. It is that they deliver video plus one mixed audio track in clips of 8 to 30 seconds, whereas GSD delivers six separable, individually editable synchronised layers at 120 seconds and beyond, at 4K and 50 frames per second, in one pass, at a materially lower price per minute. Figures use published API list pricing from the model owners and their authorised providers, converted to cost per minute as the per second rate times sixty.

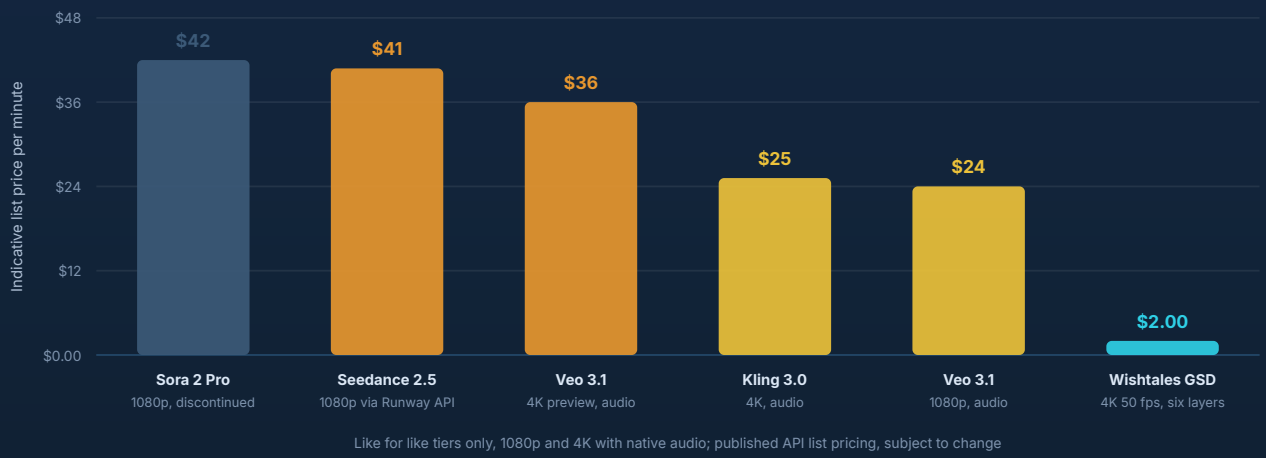


Figure 12. Cost per minute of output, like for like. At about USD 2.00 per minute of fully synchronised six layer 4K output, GSD sits 12× to 21× below the 1080p and 4K native audio tiers of the commercial frontier, from Veo 3.1 at USD 24 to the withdrawn Sora 2 Pro at USD 42. Lower tiers, silent output and 720p, are omitted because they are not like for like. The per minute figure alone does not credit GSD's layer separability, frame rate, single pass duration or included edits.

Platform	Max single pass	Resolution and rate	Output layers	Indicative list price
Wishtales GSD	120 s and beyond, no cuts	4K · 50 fps	Six separable synchronised	≈$2.00 / min, full multimodal, edits included
ByteDance Seedance 2.5	≈30 s	720p native, 1080p via providers · 24 fps	Video + mixed audio (not separable)	≈$27 to $41 / min (1080p via providers)
Google Veo 3.1	≈8 s, chained by extension	1080p native, 4K in preview · 24 fps	Video + mixed audio (not separable)	≈$24 / min (1080p), ≈$36 / min (4K)
Kuaishou Kling 3.0	≈15 s, multi shot	4K native mode · 30 fps, 60 fps option	Video + mixed audio (not separable)	≈$25 / min (4K)
Runway Gen-4.5	≈10 s	720p · 24 fps	Video; audio post processed	Silent 720p; outside the like for like price range
OpenAI Sora 2 Pro	≈20 to 25 s, six extensions	1080p · 24 fps class	Video + mixed audio	≈$42 / min (1080p); service discontinued

Pricing reflects published list pricing at the like for like tiers, 1080p and 4K with native audio, and is subject to change; lower tiers are not priced because they are not comparable. Per minute price alone is a conservative comparator because it does not credit GSD's six separable layers, frame rate, single pass duration or included edits. Section 5.2 sets the same commercial systems against thirty nine capability categories. Open weights models are excluded throughout because hosted variants can be fine tuned and their published figures are not comparable.

5.2 Full Benchmark Matrix: GSD Against the Current Frontier · Part 1 of 3 · Groups A and B

39 categories across eight capability groups, against the commercial frontier: Seedance 2.5, Veo 3.1, Kling 3.0 and Runway Gen-4.5, with the discontinued Sora 2 Pro retained as the reference for where premium pricing stood. Wishtales figures are from the internal evaluation harness on the production configuration, Blackwell B300 at FP8; competitor figures are from publicly described capabilities, published documentation and authorised provider pricing, and are approximate. Open weights models are excluded because hosted variants can be fine tuned and their published figures are not comparable.

Capability	Wishtales GSD	Seedance 2.5	Veo 3.1	Kling 3.0	Runway Gen-4.5	Sora 2 Pro (discontinued)	GSD
A · GENERATION PARADIGM AND ARCHITECTURE							
Generation paradigm	**One shot Retrieve, Optimise, Render over an explicit world graph**	Unified multimodal diffusion transformer	Diffusion transformer, native audio	Diffusion transformer, multi shot	Diffusion transformer, silent video	Diffusion transformer, native audio; discontinued	●
Synchronisation primitive	**Hierarchical graph traversal; no attention in the sync path**	Dense or learned sparse attention	Dense attention	Dense attention	Dense attention	Dense attention	●
End to end scaling in N	**O(M log N + N)**	Quadratic class	Quadratic class	Quadratic class	Quadratic class	Quadratic class	●
World model	**Explicit topological graph, quasi deterministic**	Implicit in weights	Implicit in weights	Implicit in weights	Implicit in weights	Implicit in weights	●
Synchronisation method	**By construction, one shared plan**	Joint generation, one mixed track	Joint generation, one mixed track	Joint generation, one mixed track	Post process alignment	Joint generation, one mixed track	●
Departure from cross attention	**Yes, change of complexity class**	No	No	No	No	No	●
B · OUTPUT FORMAT							
Maximum single pass duration, no cuts	**120 s and beyond**	≈30 s	≈8 s	≈15 s	≈10 s	≈20 to 25 s	●
Native resolution	**4K, sustained across 120 s+**	720p native; 1080p via provider tiers	1080p native; 4K in preview	4K native mode	720p; upscale to 4K	1080p (Pro tier)	●
Frame rate	**50 fps, sustained across 120 s+**	24 fps	24 fps	30 fps; 60 fps option reported	24 fps	24 fps class	●
Route to a 120 second asset	**One pass**	Four or more clips stitched, or multi round extension	Extension chain of fifteen or more clips	Eight or more clips	Twelve or more clips	One generation plus six extensions, each paid	●
Aspect ratios	**All common ratios**	Six ratios	16:9, 9:16	16:9, 9:16, 1:1	Six ratios	16:9, 9:16	●

● GSD leads on published capability ● Parity ● Descriptive, or no comparable competitor figure published

29 of 39
Categories in which GSD leads the field on published capability

3
Categories at parity: native resolution and frame rate, where Kling 3.0 advertises native 4K with a 60 fps option at 15 seconds, and aspect ratio coverage

7
Descriptive categories, or categories where no competitor publishes a comparable figure

5.2 Full Benchmark Matrix: GSD Against the Current Frontier · Part 2 of 3 · Groups C, D and H

Audio, synchronisation, consistency, physics and status. Competitor audio is generated with the picture but delivered as one mixed track; GSD delivers voice, music and effects as separate synchronised layers.

Capability	Wishtales GSD	Seedance 2.5	Veo 3.1	Kling 3.0	Runway Gen-4.5	Sora 2 Pro (discontinued)	GSD
C · AUDIO AND SYNCHRONISATION							
Native voice and dialogue	**Yes, separate layer**	Yes, mixed	Yes, mixed	Yes, mixed, multilingual	No; third party tools	Yes, mixed	●
Native music	**Yes, separate layer**	Yes, mixed	Yes, mixed	Yes, mixed	No	Yes, mixed	●
Native sound effects	**Yes, separate layer**	Yes, mixed	Yes, mixed	Yes, mixed	No	Yes, mixed	●
Lip sync	**By construction, multi character, across 120 s+**	Within the clip	Within the clip	Within the clip	Third party tool	Within the clip	●
Separable audio stems	**Voice, music and SFX as three layers**	Single mixed track	Single mixed track	Single mixed track	Not applicable	Single mixed track	●
On screen text as an editable layer	**Yes**	Baked into pixels	Baked into pixels	Baked into pixels	Baked into pixels	Baked into pixels	●
Independently editable synchronised layers	**6**	2	2	2	1	2	●
D · CONSISTENCY AND PHYSICS							
Character identity across the runtime	**Topological invariant, quasi deterministic across 120 s+**	Up to fifty reference assets per clip (images, video, audio)	Up to three reference images	Elements: tracked characters and reference images	Single reference image	Reference image; world state within one generation	●
Prompt constraint adherence	**99.89% (internal harness, 10,000 outputs)**	Not published	Not published	Not published	Not published	Not published	○
Prompt to physics adherence	**98.99% (internal harness)**	Not published	Not published	Not published	Not published	Not published	○
Physics validation loop	**Parallel video reasoning layer, per frame**	None described	None described	None described	None described	None described	●
Continuity across scenes without regeneration	**Persistent creative state**	References and extension	Scene extension	Multi shot in one call	Not native	Extensions, each a new generation	●
Safety boundary	**Ontological omission; unpopulated domains cannot be generated**	Post generation filter	Post generation filter	Post generation filter	Post generation filter	Post generation filter	●
H · STATUS AND RESEARCH POSTURE							
Model status	**Production on Blackwell B300; 120 s+ at 4K 50 fps in ≈240 s; six patent families filed**	Active	Active; no successor announced	Active	Active; flagship silent	Discontinued; consumer app closed, API sunset announced	○
Ownership of the generative core	**Wishtales owned and patented**	ByteDance; resold through US platforms	Google	Kuaishou; resold through US platforms	Runway; resells Seedance 2.5, Kling 3.0 and Veo 3.1	OpenAI	○

● GSD leads on published capability ● Parity ○ Descriptive, or no comparable competitor figure published

5.2 Full Benchmark Matrix: GSD Against the Current Frontier · Part 3 of 3 · Groups E, F and G

Editability, speed and economics. Price per minute is the per second list rate times sixty at each system's like for like tier, 1080p or 4K with native audio, with provider resale rates shown where the model owner sells a resolution only through providers; lower tiers and silent output are not priced because they are not comparable. The 120 second asset row multiplies that by two and, for clip systems, notes that assembly across clips is a further cost. Hosted median generation times are thirty day medians reported by hosting platforms and vary with queue depth.

Capability	Wishtales GSD	Seedance 2.5	Veo 3.1	Kling 3.0	Runway Gen-4.5	Sora 2 Pro (discontinued)	GSD
E · EDITABILITY AND CREATIVE CONTROL							
Layer selective edits	**Giga Papaya: any layer, any second**	Timestamp and region targeted edits within the clip; mixed file re rendered	Regenerate, extend, or insert and remove; whole clip	Regenerate the clip	Aleph video to video, whole clip	Targeted edit endpoint; whole clip re rendered	●
Edit quantisation	**One second, per layer**	Clip segment	Clip	Clip	Clip	Clip	●
Cost of an edit	**Scales with the edit; 65 to 90% below re render**	Re render of the clip as one mixed file	Full clip regeneration	Full clip regeneration	Full clip regeneration	Full clip regeneration	●
Duration changing edits reflow other layers	**Yes, automatic**	No	No	No	No	No	●
Propagation of an edit to later scenes	**Yes, on confirmation**	No	No	No	No	No	●
Edits included in the price	**Effectively unlimited**	Metered per second	Metered per second	Metered per second	Metered per second	Metered per second	●
F · SPEED							
Wall clock for 15 s of output	**≈30 s, six layers, 4K 50 fps**	≈3.5 to 4 min per 8 s clip (hosted median)	Under 2 min per 8 s clip (hosted median)	≈1.5 min per 8 s clip (hosted median)	Not published	≈30 to 90 s per 5 to 10 s clip (aggregator report)	●
Wall clock for 120 s of output	**≈240 s, one pass; 154 s in ≈300 s**	Multiple clips plus assembly	Multiple clips plus assembly	Multiple clips plus assembly	Multiple clips plus assembly	Seven generations plus assembly	●
Output seconds per wall clock second	**≈0.5, sustained**	≈0.03 to 0.04	≈0.07	≈0.09	Not published	≈0.1 to 0.2 (reported)	●
G · COST AND PRODUCTION ECONOMICS							
List price per minute, like for like tier with audio	**≈$2.00; 4K 50 fps, six layers**	≈$27 to $41 (1080p via providers)	≈$24 (1080p); ≈$36 (4K preview)	≈$25 (4K)	Silent 720p; outside the like for like range	≈$42 (1080p)	●
Cost of a 120 second asset with audio	**≈$4.00, edits included**	≈$54 to $82 at 1080p, plus assembly	≈$48 to $72 plus chaining	≈$50 at 4K plus assembly	Silent; not a like for like comparator	≈$84 at 1080p	●
GPU memory for 30 s of output	**≈12 GB**	Not disclosed	Not disclosed	Not disclosed	Not disclosed	Not disclosed	○
Gross margin posture	**50% pure GSD; 35 to 40% with unlimited edits**	Not disclosed	Not disclosed	Not disclosed	Not disclosed	Not disclosed; withdrawn on cost grounds	○

● GSD leads on published capability ● Parity ○ Descriptive, or no comparable competitor figure published

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; other foundational model companies require USD 50 million to USD 100 million of investment at the pre seed and seed stage, before any production ready output exists. The current version of GSD and Giga Papaya can be trained to production standard and brought to market direct to consumer for between USD 1.5 million and USD 5 million in total, an order of magnitude less at every representative endpoint. Future versions, scale and go to market will require more capital; the current version can be brought to market within that envelope.

$1.5 to 5M
Can bring the current GSD and Giga Papaya to market, direct to consumer

$50 to 100M+
Other foundational models, before production ready output

$2.00
Customer price per 60 s, edits included

50%
Blended gross margin on pure GSD output, day one

6.1 Profitable Unit Economics From Day One

The customer price is approximately USD 2.00 per 60 second multimodal output, for 4K at 50 frames per second with all six layers. Because the architecture removes the quadratic cost driver, pure GSD generation carries a blended gross margin of 50 per cent at that price from day one, and significant positive gross margins at the pod level under conservative operating assumptions. Giga Papaya edits are included in the same price rather than metered; since each edit is a quantised operation on a bounded slice of the timeline rather than a re render, the blended margin including an effectively unlimited edit allowance settles between 35 and 40 per cent depending on edit volume, in an industry where well funded competitors operate at negative gross margin and where the most prominent product has been withdrawn on cost grounds. 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.

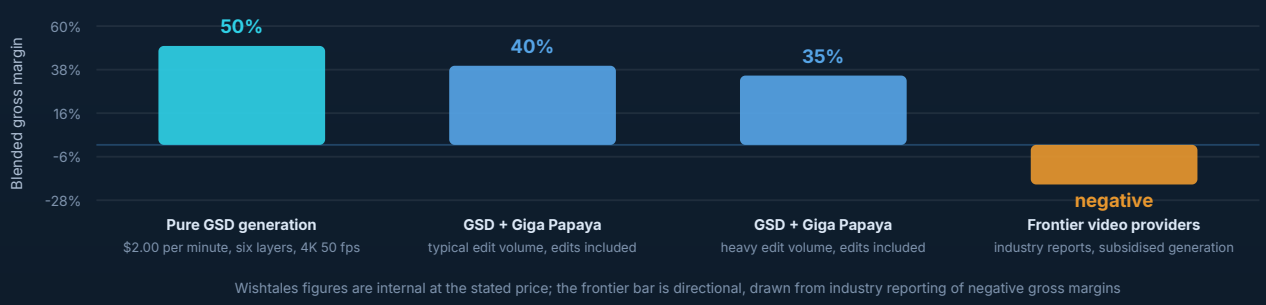


Figure 13. Blended gross margin at the USD 2.00 per minute price. Pure generation carries 50 per cent; including an unlimited Giga Papaya edit allowance the margin settles at 35 to 40 per cent by edit volume.

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.

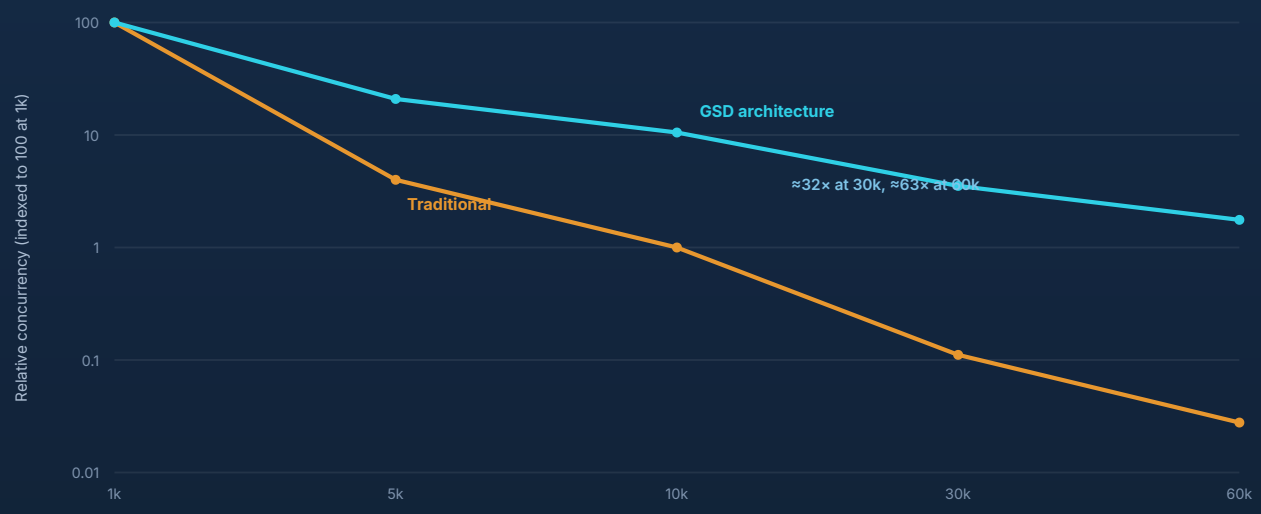


Figure 14. Relative sustainable concurrency versus sequence length. The capacity ratio widens as Θ(N): about 32× at 30k tokens and about 63× at 60k, growing without bound thereafter. A 120 second 4K timeline sits toward the right of this axis.

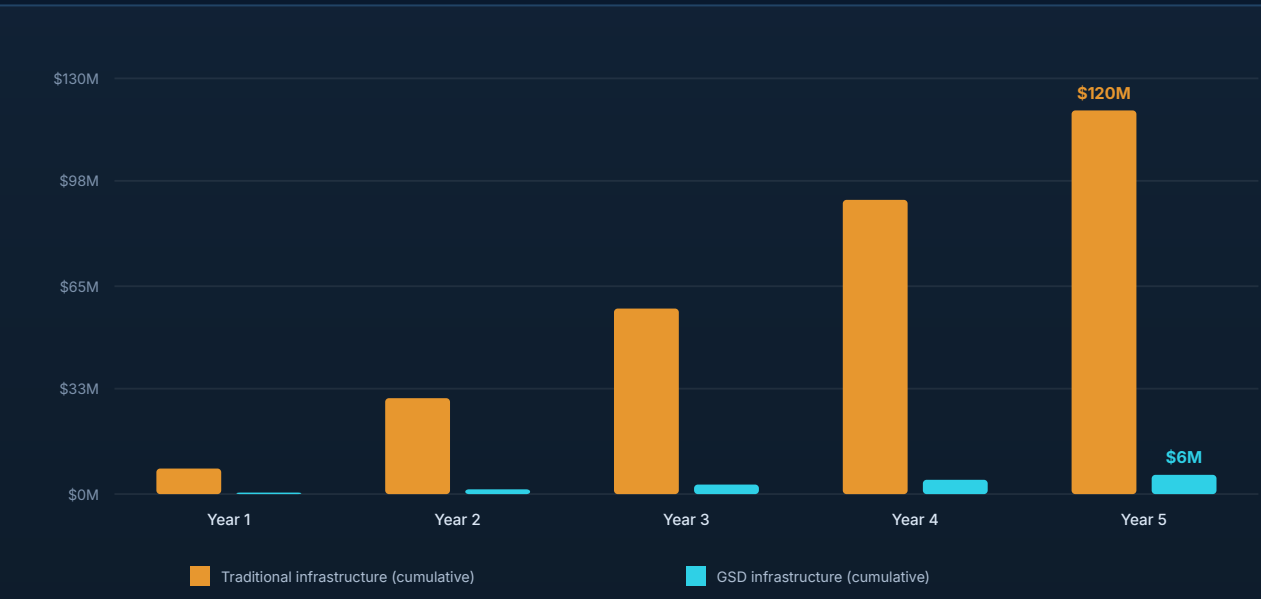


Figure 15. Illustrative five year cumulative infrastructure cost for 100,000 concurrent users generating 60 minutes per month, under the stated cost model. Cumulative advantage approaches 20×.

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. Production targets Blackwell class accelerators, on which the 4K, 50 frames per second, 120 second configuration is delivered.

07 What Creators Need: The Four Cs

Across creators, agencies, studios and enterprises, the requirements of a video system reduce to four: it must be convenient, its output must be consistent, it must give creative control, and it must be affordable. The four are usually traded against each other. GSD is designed so that all four follow from the same architectural decision.

Convenience C

One prompt, one pass, one finished asset. Video, voice, lip sync, sound effects, music and on screen text arrive together and in sync, in about twice the running time of the output. No storyboard of eight second clips, no separate voice tool, no alignment pass, no stitching, and no queue of retries to reach a usable take.

Proof: 120 s+ in ≈240 s on Blackwell B300 · six layers in one pass · a fifteen second asset in about thirty seconds

Consistency C

Identity is a topological invariant of the world graph, not a surface feature of a frame, so the same character keeps the same face, voice and physical behaviour from the first second to the last, across scenes and across edits. Physics is held to the same standard on every layer, and a persistent creative state carries characters and sets from one project to the next.

Proof: 98.99% prompt to physics adherence · 99.89% constraint adherence · quasi deterministic identity across 120 s+

Creative Control C

Every layer is separable and every second is addressable. Giga Papaya lets a creator re voice a line, swap the music, change a costume or extend a shot and re render only the implicated layer for only the implicated seconds, with changes propagated to later scenes on confirmation and every other layer reflowed automatically when a duration changes. Multi channel creative control is native, not bolted on.

Proof: six independently editable layers · one second edit quantisation · effectively unlimited edits inside the price

Cost C

USD 2.00 per minute for the complete six layer 4K asset at 50 frames per second, with edits included and a 50 per cent margin already inside the price, against USD 24 to USD 42 per minute for a single video layer with a mixed audio track on the premium 1080p and 4K tiers of the commercial frontier. The price is sustainable because it is not subsidised.

Proof: 12× to 21× below premium native audio pricing · 50% margin on pure generation · 35 to 40% with unlimited edits

Requirement	GSD, one platform	Frontier clip model plus tool stack
Convenience	**One prompt, one pass, one asset, about 2× real time**	Clip generation, separate voice, music and effects tools, alignment and stitching; minutes per clip plus assembly
Consistency	**Quasi deterministic identity and physics across 120 s+**	Reference images per clip; drift at every clip boundary
Creative control	**Any layer, any second, propagated and reflowed**	Regenerate the whole clip for any change; audio is one mixed track
Cost	**$2.00 per minute, edits included, positive margin**	$24 to $42 per minute for video with mixed audio at the premium tiers, plus tool subscriptions, metered edits and rejected takes

08 Applications Across Creative Domains

The applications of GSD share one economic fact: fully synchronised multimodal output at roughly one twelfth to one twentieth the cost of premium native audio video, with every layer separable and editable, at 120 seconds and beyond per pass.

8.1 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 at 4K and 50 frames per second, GSD is more than an order of magnitude more economical than premium APIs that price video with a single mixed audio track at USD 24 or more per minute, and it replaces the four or five separate tools a creator would otherwise pay for. A two minute product video, explainer or short narrative arrives as one asset in about four to five minutes, with unlimited edits included. This brings studio grade multimodal production within reach of individual creators.

8.2 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.

8.3 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. Because voice is a separable layer, localising a finished asset is a Giga Papaya operation on the voice head rather than a regeneration. The figure is a modelled scenario rather than a committed third party result.

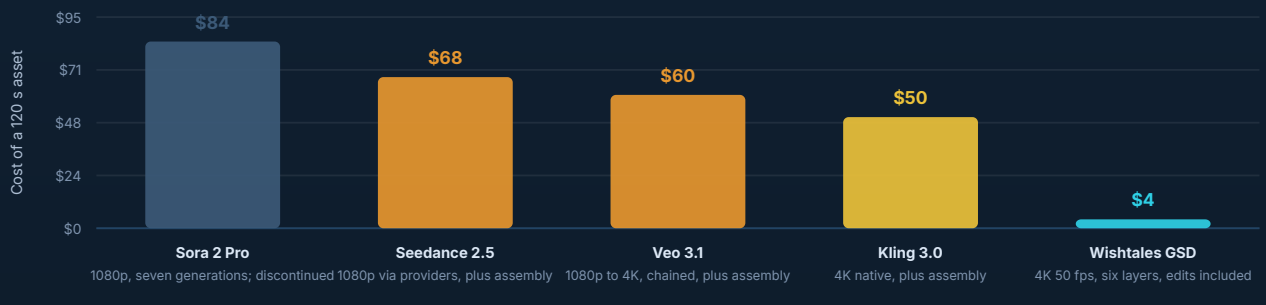


Figure 16. What a two minute asset costs at list price, like for like. GSD delivers the complete six layer 4K asset for about USD 4, with edits included; the withdrawn Sora 2 Pro would have charged about USD 84 for one 1080p video with a mixed track.

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.

8.4 The Back End of the Creator Platforms: GSD and Giga Papaya by API

Every platform in Figure 1 already integrates a third party generative core through a per second API, so the integration GSD requires is the one they operate today; what changes is what the call returns. A generate call returns six synchronised, separable layers and the blueprint that produced them, at 120 seconds and beyond, instead of one mixed file of eight to thirty seconds. An edit call re renders one layer for a handful of seconds and reports the reflow, instead of a fresh clip. The consumer facing surfaces of the industry, Runway, Luma, Pika, ElevenLabs and the platforms that route to Chinese cores, can therefore adopt GSD and Giga Papaya as their back end without changing their front end, and enterprise integrations of exactly this form are under way.

GENERATE: ONE CALL, ONE ASSET

```
POST /v1/generate
{
  "prompt": "ad commercial for a glow in the dark waterproof
smart shoe",
  "duration_s": 120,  "resolution": "4K",  "fps": 50,
  "layers": ["video", "voice", "lip_sync", "sfx", "music",
"on_screen_text"],
  "references": [{"type": "image", "url": "..."},
                 {"type": "voice_clone", "id": "..."}]
}
returns { "asset_id", "blueprint_id",
          "layers": {six synchronised stems},
          "timeline": {per second index} }
```

EDIT: ONE LAYER, A FEW SECONDS

```
POST /v1/edit
{
  "asset_id": "...",
  "layer": "voice",
  "seconds": [47, 53],
  "instruction": "re voice the line, warmer and slower",
  "propagate": true
}
returns { "slice": {re rendered seconds, one layer},
          "reflow": {timing shifts on other layers},
          "later_occurrences": [...] }
```

Platform layer	Generative core today	What the same integration returns with GSD and Giga Papaya
Runway	Gen-4.5 at 720p silent; resells Seedance 2.5, Kling 3.0 and Veo 3.1 by the second	**120 s+ six layer 4K 50 fps assets in one call; per second edits inside its own timeline**
ElevenLabs	Resells Veo, Kling and Seedance beneath its voice stack	**Picture, lip sync, effects and music generated in sync with its voices; voice stays a separable layer**
Luma and Pika	Short clip models and agentic interfaces over existing cores	**A long form, physics faithful core beneath the same interface, at a wholesale rate that leaves margin on both sides**
Higgsfield, Krea, Freepik, OpenArt	Interfaces over Chinese clip models	**A domestically owned, patented core with no clip ceiling and edits included in the price**
Creative suites hosting partner models	Partner clip models selected per shot	**One partner model that returns the whole piece, every layer editable in the host timeline**

Creator and social apps
Topic to finished post: video, voice, captions and music in one asset, edited by the second before publishing.

Advertising and performance creative
Hundreds of two minute variants from one brief; swap a line, a product shot or a market with an edit call.

Retail and product video
Catalogue to product videos at 4K, with the same product identity held across every shot and every variant.

Localisation and dubbing
Re voice the voice layer per market; lip sync, music and effects reflow automatically.

Education and explainers
Script to narrated, animated lesson; Spark compiles diagrams and equations in sync with the narration.

Games, previs and cinematics
Storyboard to animatic to cinematic from one blueprint, with physics held from the first frame to the last.

Broadcast, streaming and sport
Recaps, promos and localised bumpers generated to length, at broadcast frame rates, minutes after the event.

Enterprise communications
Training, onboarding and product updates regenerated from a persistent creative state as the content changes.

Agents
Autonomous agents call generate and edit over the protocol layer of Section 10 to produce and revise video without a human in the loop.

Field names above illustrate the shape of the two calls; the production schema, authentication and rate model are shared under agreement. Section 10.7 sets out the wholesale economics for a platform adopting GSD as its core.

09 Intellectual Property: Six Patent Families

Wishtales has filed six United States provisional patent applications, with the corresponding foreign filing permissions sought in India, covering the generation core, the editing engine, the world model layer and the three creator surfaces built on them. Together they protect the fourth pillar of Section 1.2 end to end, from the world graph to the timeline a creator touches. Each family is summarised below in three registers: the technical claim in one sentence, the same claim in plain words, and what it means for the person using the product, followed by why it sits ahead of what exists in the market today. The specifications, claims and drawings are not reproduced here.

P1 Generative Synchronous Diffusion (GSD)

Generation core

TECHNICALLY	The world's first one shot long form video generation model with quasi deterministic character and physics consistency, using proprietary Topological ML and World Graph techniques of topology and graph theory to retrieve, optimise and render multimodal video with voice, sound effects, background music and lip sync all at once and all in sync in a single pass, through multi head renders with multi channel creative control, at orders of magnitude less compute and cost than cross attention or transformer diffusion video models.
IN PLAIN WORDS	Describe the video once and get the whole thing, picture and every sound, in one go, with the same people and the same physics from the first second to the last.
USER BENEFIT	**120 seconds and beyond of 4K video at 50 frames per second, in about twice its running time, at USD 2 per minute.** One tool replaces a video model, a voice tool, a music tool, an effects tool and an editor.
WHY IT IS AHEAD	Every competing model, American or Chinese, is an attention based diffusion transformer with an 8 to 30 second ceiling and a single mixed audio track. GSD changes the complexity class, so the advantage widens with every additional second and every additional layer.

P2 Giga Papaya

Editing engine

TECHNICALLY	A localised, layer selective multimodal video editing engine that uses temporal localisation, quantised to every second of every layer, to confine an edit to a slice of the full length video and re render only the implicated modality for only the chosen seconds in a quasi deterministic manner, propagating the change through later scenes and automatically readjusting every other layer when a duration changes.
IN PLAIN WORDS	Change one thing, in one place, and only that thing changes; the rest of the video stays exactly as it was, and the change follows the character into later scenes if you want it to.
USER BENEFIT	**Unlimited edits inside the price.** An edit costs seconds of one layer, not a re render of the asset, which is what makes creative control affordable rather than metered.
WHY IT IS AHEAD	Every other system regenerates the whole clip for any change, and none exposes separable voice, music, effects or text layers to edit in the first place.

P3 Synthesis World Model Layer

Persistent world state

TECHNICALLY	A real time multimodal perception, reasoning and synthesis layer operating through an Abstract, Reason, Synthesise pipeline that abstracts sight, sound, motion and spatial input directly into a persistent, modality independent representation of the world, uses topological signatures to resolve different observations of the same entity or event to a shared world state identity, and synthesises that state into synchronised media, interactive virtual environments, machine interfaces and physical AI control without a compulsory text layer.
IN PLAIN WORDS	The system keeps one living model of the world and everything in it, so what it sees, hears and generates all refer to the same things, without translating everything into words first.
USER BENEFIT	**Persistent characters, sets and continuity across projects.** Edits become state changes, a creator's library of styles and characters is reusable creative state, and the same core can drive interactive and physical systems.
WHY IT IS AHEAD	Prevailing systems route perception through text and hold no persistent world state, so each generation starts from nothing and continuity has to be re described every time.

P4 Supercut Mobile

Mobile creator surface

TECHNICALLY	A cross platform multimodal generative AI video editor for mobile that routes spoken or typed edit intent through an editor module registry to a temporal slice of an individual layer on a multi track synchronised timeline, with an on device timeline to filtergraph compiler rendering the final video on the handset and one synchronised project representation shared across Android and iOS.
IN PLAIN WORDS	Speak or type any edit, and it lands on the right layer at the right second of a full length AI video, on your phone, and plays immediately.
USER BENEFIT	**Edit any layer at any second from a phone, with no desktop, no upload and no waiting.** Start on one device and finish on another with every change already there.
WHY IT IS AHEAD	Mobile video tools today either upload to the cloud and wait, or edit a single flat track; none addresses individual synchronised layers of an AI generated video on the device.

P5 Compose Web

Web authoring surface

TECHNICALLY	A multimodal document authoring system in which people, teams and AI co write one document in words, pictures, video and sound, with the AI as a first class editor inside the document rather than a side panel beside it, every contribution from every person and every model assembled quasi deterministically, and the single source document transformed into any output format required.
IN PLAIN WORDS	Write it once, with your team and the AI, and turn it into any format you need: a presentation for the boardroom, a narrated audiobook, a video for your channel, an illustrated storybook.
USER BENEFIT	**Never build the same content twice, and nothing is silently lost between formats.** One source document becomes every format of your choice, with a GSD video among them.
WHY IT IS AHEAD	Existing editors bolt AI on as an assistant beside the page and export one format; they do not assemble multi author, multi model, multimodal contributions into one governed source.

P6 Wishtales Spark

Explainer engine

TECHNICALLY	A semi deterministic explainer video engine that compiles animation from structured mathematical and diagrammatic graph representations, animates a small set of still images and graphics rather than synthesising every frame, and aligns every animation event to the measured duration of a generated voiceover.
IN PLAIN WORDS	Give it a topic and a voice and it draws the explanation for you, with diagrams that build and move in step with the narration, finished in minutes.
USER BENEFIT	**A narrated explainer video at roughly one tenth of the graphics processing of a comparable AI video.** Correct diagrams, on time with the voice, for any subject and any language.
WHY IT IS AHEAD	Frame synthesis models cannot draw a correct diagram or hold a formula steady; Spark compiles the explanation rather than hallucinating it, at a fraction of the compute.

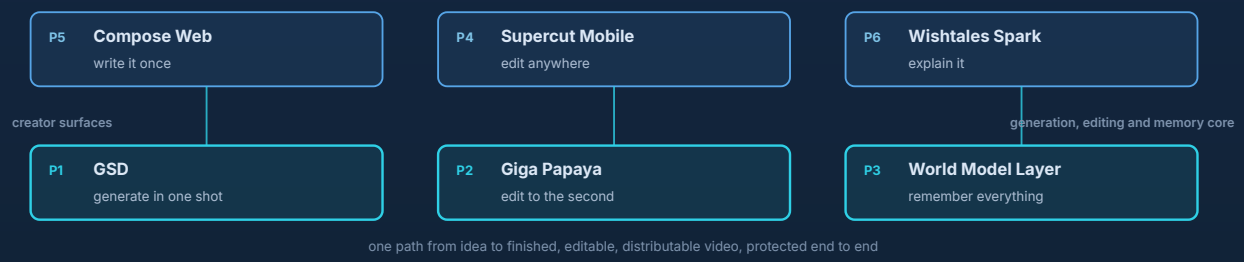


Figure 17. The six families as one path from idea to video: three creator surfaces above, and the generation, editing and memory core beneath them.

9.1 How the Six Families Fit Together

Together the six families protect the full path from idea to finished, editable, distributable video: the generation core (P1), the editing engine that makes its layers addressable (P2), the persistent world model that gives it continuity (P3), and the mobile (P4), web (P5) and explainer (P6) surfaces through which creators reach it. The user facing consequence is the Four Cs of Section 7; the investor facing consequence is that the moat is not a model checkpoint that a competitor can retrain past, but a change of complexity class and the tooling that only that change makes possible.

Family	Protects	Answers	Measured by
P1	One shot, six layer, long form generation	Convenience, Consistency, Cost	120 s+ · 4K 50 fps · ≈67× · $2.00 per minute
P2	Layer selective, second quantised editing	Creative Control, Cost	1 s quantisation · 65 to 90% edit compute saving
P3	Persistent, modality independent world state	Consistency	Identity held across scenes, projects and edits
P4	On device, multi layer mobile editing	Convenience, Creative Control	Any layer, any second, on the handset
P5	Multimodal, multi author document authoring	Convenience	One source, every format
P6	Compiled explainer animation synced to voice	Convenience, Cost	About one tenth of comparable AI video compute

Patent pending. Filing details, inventorship and claim language are matters of record with counsel and are not part of this overview.

10 GSD Versatility: Multiple Avenues of Manifestation

GSD deploys across infrastructure and application domains as an integration layer, an API service, or a standalone platform, and, for the platforms that today resell third party clip models, as the generative core beneath their own interfaces.

10.1 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.

10.2 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.

10.3 Creative Platform, Studio and Enterprise Integration

Domain	Today	With GSD	Projected impact
Creative platform API	Fragmented per task models, increasingly resold from third parties	**Single synchronised endpoint, owned core**	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; edits by the second**	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.

10.4 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.

10.5 Premium Streaming and Enterprise Video

The most demanding deployment is premium streaming and enterprise video, 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 delivers 4K at 50 frames per second for 120 seconds and beyond in a single pass, with a parallel video reasoning layer that validates physics and corrects on screen text. This is the configuration targeted for premium streaming and enterprise video delivery.

10.6 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 quasi deterministic identifiers enables instant rollback to previous creative decisions without losing subsequent work, which moves production management from reactive troubleshooting to proactive optimisation.

10.7 A Foundational Core for the Platforms That Resell Clip Models

The creator platforms that today place an interface over Seedance or Kling pay a per second wholesale rate to a foreign model owner, add their own margin, and inherit the clip ceiling, the mixed audio track and the negative economics beneath them. GSD offers the same platforms a domestically owned core that produces six separable layers at 120 seconds and beyond, at a wholesale rate that leaves room for a margin on both sides, because the USD 2.00 retail price already carries 50 per cent. The integration surface is the pair of calls set out in Section 8.4, exposed through the synchronised endpoint of Section 10.3.

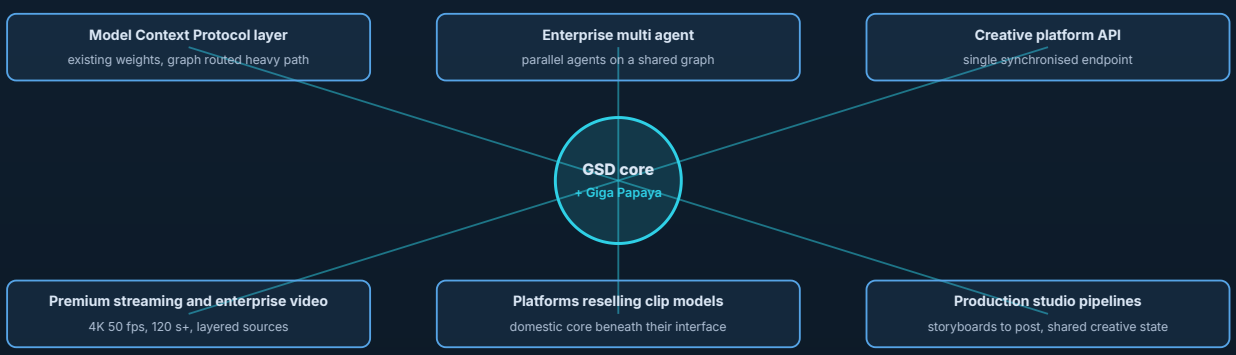


Figure 18. Avenues of manifestation. One core, one editing engine, six routes to market, each reached through the same synchronised endpoint.

11 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 model size and compute are the primary path to capability. The largest compute budgets in the industry have produced eight to thirty second clips at a loss, and the most prominent of them has been withdrawn. The trade off in GSD is explicit and deliberate: it is engineered as a synthesis engine, not a general purpose reasoning engine, and its domain boundaries are verifiable by ontology and graph population, which is at once a capability boundary and a safety property.

11.1 Why This Is a Paradigm Shift Rather Than an Increment

Three tests separate a new pillar from a better clip model, and GSD passes all three. The complexity class changes, $O(M \log N + N)$ against $\Theta(L \cdot M^2 \cdot N^2)$, which is why the horizon moved from thirty seconds to 120 seconds and beyond in one step. The output changes in kind, six separable synchronised layers and an editable blueprint rather than one mixed file, which is what makes per second editing possible at all. And the economics invert, 50 per cent gross margin at USD 2.00 per minute where the incumbents lose money at ten to twenty times the price and the most heavily funded of them has been withdrawn. The six patent families of Section 9 protect each stage of that mechanism; the measurements of Section 4, 120 seconds and beyond at 4K and 50 frames per second in about 240 seconds on Blackwell B300 with 98.99 per cent physics adherence across the full timeline, are the evidence that it works at production scale.

11.2 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 quasi 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.

11.3 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 protected as its own patent family (Section 9) and is described separately; it 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 customer outcome; realised efficiency depends on hardware, and the integration scenarios in Section 10 are modelled rather than committed third party results. Competitor figures are drawn from public documentation and provider pricing and will move. Single pass outputs well beyond 154 seconds are a matter of accelerator memory headroom rather than of architecture and are not claimed here. Given positive unit economics, the principal risk is not the economics but engineering throughput, Blackwell capacity and time to market.

12 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 difference in capital: the current version of GSD and Giga Papaya can be brought to market direct to consumer for between USD 1.5 million and USD 5 million, with future versions requiring more, against USD 50 million to USD 100 million for other foundational models before production ready output, an approximately 67× realised inference efficiency gain at fixed quality, a 99.89% prompt constraint adherence rate and a 98.99% prompt to physics adherence rate, with physics coherently represented across all modes. At production scale the result is a single pass that produces 120 seconds and beyond of six layer 4K video at 50 frames per second in about twice its running time, editable to the second, at USD 2.00 per minute with a 50 per cent margin inside the price.

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, and the withdrawal of the field's most prominent product on cost grounds has made the point for us. GSD is constructed to be one such architecture: the fastest, the least expensive and, across the thirty nine categories of Section 5, the most capable, and it is offered as a foundational model and a fourth architectural pillar for the next phase of multimodal AI.

Five Numbers That Define the Gap

Measure	Wishtales GSD	Best commercial alternative	Gap
Single pass duration, no cuts	**120 s and beyond; 154 s demonstrated**	30 s (Seedance 2.5)	4× or more
Native resolution and rate, sustained	**4K at 50 fps across the whole pass**	4K at up to 60 fps for 15 s (Kling 3.0); 720p at 24 fps for 30 s (Seedance 2.5)	Same fidelity, eight times the length
Speed, output seconds per wall clock second	**≈0.5 on Blackwell B300, six layers**	≈0.03 to 0.09 hosted medians, one layer	5× to 17×
Price per minute with audio, like for like	**USD 2.00, six layers, edits included**	USD 24 to USD 42 at 1080p and 4K with native audio (Veo 3.1 to Sora 2 Pro)	12× to 21×
Independently editable synchronised layers	**6, quantised to one second**	2 (video plus one mixed track), whole clip	Different in kind

≈2× real time
Generation speed at six layers and 4K

98.99%
Prompt to physics adherence

50%
Gross margin on pure GSD output

6
Patent families filed

> 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, and why a 120 second, six layer asset is a natural output of the architecture rather than a stretch of it.

§ 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.
7. OpenAI Help Center. What to know about the Sora discontinuation; and contemporaneous reporting by CNN Business, NBC News and TechCrunch on the closure of the Sora app and API and the dissolution of the Disney agreement.
8. Google Cloud, Vertex AI Veo 3.1 pricing and model documentation, including per second rates with and without audio at 1080p and 4K, clip duration and scene extension.
9. ByteDance Seedance 2.5 documentation and pricing as published by BytePlus ModelArk (native 480p and 720p at 24 frames per second, 4 to 30 seconds, published five second example rates) and by authorised providers offering 1080p tiers (Runway developer API, Atlas Cloud, EvoLink, Replicate), including per second rates.
10. Kuaishou Kling 3.0 documentation and pricing as published by Kling AI and authorised API providers (fal, Replicate), including the native 4K mode and clip duration.
11. Runway developer API pricing and model documentation for Gen-4.5 (12 credits per second at USD 0.01 per credit, 720p, 5 or 10 seconds, no native audio), and Runway API and plan documentation listing Seedance 2.5 at 30 credits per second (720p) and 68 credits per second (1080p), Kling 3.0 and Veo 3.1 as resold third party models.
12. OpenAI developer platform, Sora 2 and Sora 2 Pro model pages and pricing (USD 0.10 per second for Sora 2 at 720p; USD 0.30, 0.50 and 0.70 per second for Sora 2 Pro at 720p, 1024p and 1080p), the extensions endpoint, and the discontinuation notices for the consumer app and the API.
13. Vaswani, A., et al. (2017). Attention Is All You Need. NeurIPS.
14. Peebles, W., & Xie, S. (2023). Scalable Diffusion Models with Transformers. ICCV.
15. Gu, A., & Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.
16. ElevenLabs, Introducing ElevenLabs Image & Video, listing Veo, Sora, Kling and Seedance among the video models offered inside the product.
17. Higgsfield, How Long Does AI Video Generation Take, thirty day median generation times by model.

AUTHORS

Animikh Roy · CEO, CTO & Product Architect
Neel Roy · Co-Author · Founding Team, Head of Growth

ENTITY

Wishtales AI Inc.
Delaware C-Corporation · India subsidiary

CONTACT

admin@wishtales.ai
Palo Alto, California

© Wishtales AI Inc. All rights reserved. · Public Technical Overview, Version 2026 · Cleared for general distribution. Proprietary mechanisms intentionally omitted. GSD, Giga Papaya, Supercut, Compose and Spark are products of Wishtales AI Inc.; patents pending. All other product names are the property of their respective owners and are used for identification only. Competitor capabilities and pricing are drawn from public sources and are subject to change.