

SPATIAL AI

Spatial AI: A Living Gaussian-Splat Field That Predicts, at the Edge

A concrete definition — an explicit field of Gaussian splats, coupled to grounded dynamics, accounted in joules and run inside a robot's power budget rather than a datacenter's.

David Jean Charlot, PhD

Dean of Physical AI · The Charlot Lab, Institute for Physical AI @ BMI

Correspondence: contact@physicalai-bmi.org · physicalai-bmi.org

Interactive companion: physicalai-bmi.org/research/charlot-lab#topic-spatial

ABSTRACT. "Spatial AI" is the phrase the field reached for after "spatial computing," and it remains a buzzword in search of a definition. This report proposes a concrete one and defends it. A spatial model should hold the world as an explicit field of points, and Gaussian splatting is the representation that captures real geometry and appearance while remaining inspectable and editable. It should also be a world model that knows how that field moves, meaning grounded dynamics rather than a frozen scan. Spatial AI is the composition of the two: a living splat field that predicts. Most current systems build this in the datacenter. The constraint studied here is the opposite one, which the lab labels Swap-2C — the model runs at the edge, on the low-power device that lives in the world it is modeling. Under that constraint the representation stays explicit, the dynamics stay grounded, and the whole computation is accounted in joules on the lab's MathGround substrate. Section 2 states the review method. Sections 3 and 4 develop the two halves of the definition. Section 5 sets out the edge constraint and the state of on-device splatting. Section 6 describes the energy accounting. Sections 7 and 8 give the position and conclusion. The report reports no new experimental measurements.

1. Introduction

Every few years the field of embodied and perceptual computing adopts a term before it agrees on its meaning. "Spatial computing" named a set of interaction and display technologies for placing content in three dimensions. "Spatial AI" arrived after it, and is now applied to almost anything that touches geometry: depth estimation, scene reconstruction, occupancy prediction, headset passthrough, robot mapping. The breadth is a symptom. A term that

covers everything constrains nothing, and a research programme cannot be built on a word that has not been given edges.

This report gives the word edges. The claim is that a system deserves the name Spatial AI when it satisfies two requirements at once. The first is representational: it holds the world as an *explicit field* of primitives with real geometry and appearance, not as an opaque latent vector or a rendered image. The second is dynamical: it is a *world model*, in the sense that it knows how that field evolves in time and can predict the next state of the scene rather than only describe the present one^[1]. A frozen scan meets the first requirement and fails the second. A video-prediction network that hallucinates future frames may meet the second and fail the first, because its state is not an inspectable field. Spatial AI is the intersection: an explicit field that carries its own dynamics, a living splat field that predicts.

The representation the lab adopts for the field is 3D Gaussian splatting^[2], for reasons developed in Section 3. The dynamics are treated in Section 4. The distinguishing constraint of this work, and the reason it is a lab programme rather than a restatement of existing systems, is stated in Section 5: the model must run at the edge, on the embodied device, within a power budget measured in watts rather than kilowatts. Section 6 makes that constraint quantitative by accounting the computation in joules.

2. Scope and method

This is a review and a research position. It surveys published representations and world models for spatial scenes, draws a definition from that literature, and states where the lab's programme sits relative to it. It reports no original experiments and presents no benchmark numbers of its own. Where it cites a quantity, that quantity is attributed to a verified source. Several of the systems discussed are recent, and the report marks explicitly where a capability is demonstrated in published work, where it is an early prototype, and where it is a conjecture or a design target of the lab rather than a measured result. The MathGround substrate and the Swap-2C edge constraint referenced throughout are internal lab constructs; claims about them are stated as design intent, not as validated performance. The principal limitation of the review is that on-device, dynamic splatting is an area of active development, and the balance between fidelity, speed, and energy is not yet settled in the literature.

3. The explicit field: Gaussian splats as the representation

The first half of the definition asks for an explicit field. The candidates are meshes, voxels, point clouds, neural radiance fields, and Gaussian splats. Neural radiance fields represent a scene as the weights of a small network queried along camera rays^[3]; they capture appearance well but store the scene implicitly, so the geometry is not directly addressable and rendering requires many network evaluations per pixel. 3D Gaussian splatting takes the opposite stance^[2]. The scene is a set of anisotropic Gaussians, each an explicit primitive with a position, a covariance, an opacity, and a view-dependent colour. Each primitive is a small ellipsoid in space,

$$G_i(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right), \quad \Sigma_i = R_i S_i S_i^\top R_i^\top,$$

where the covariance Σ_i is factored into a rotation R_i and a scale S_i so it stays positive semi-definite during optimization. An image is formed by projecting the Gaussians to the camera and compositing them front to back with the standard over operator,

$$C(\mathbf{p}) = \sum_i c_i \alpha_i \prod_{j < i} (1 - \alpha_j),$$

where α_i is the projected Gaussian's opacity at pixel $\mathbf{p}$ and c_i its colour. Because the compositing is differentiable, the whole field is fit to a set of photographs by gradient descent, and because the primitives are explicit the fit renders in real time by rasterization rather than ray marching^[2].

Explicitness is the property that matters for Spatial AI, and it has three consequences the lab treats as first-class. The field is *inspectable*: a splat has coordinates, so one can ask what is at a location and get an answer without decoding a latent. It is *editable*: individual primitives can be moved, removed, or relit, which is what makes dynamics expressible as operations on the field rather than as a re-run of a black box. And it is *metric*: the primitives live in the same space as the robot, so the representation is directly usable for planning and control. A latent world model can be accurate and still fail all three tests. This is the reason the lab fixes the representation to an explicit field and, within that class, to Gaussian splats, which currently give the best combination of appearance fidelity and real-time explicit rendering.

4. From a scan to a world model: grounded dynamics that predict

A splat field on its own is a scan. It reconstructs the scene as it was when the photographs were taken and freezes it. The second half of the definition asks for more: the field must carry dynamics, so that given the state at time t it can produce the state at $t + \Delta t$. Two lines of published work show that this is feasible on the explicit representation. Dynamic and four-dimensional Gaussian splatting extend the static field with per-primitive motion, either by tracking each Gaussian through time as a persistent particle^[4] or by attaching a deformation field that carries the primitives forward, giving real-time rendering of a moving scene^[5]. In the simplest grounded form the field advects,

$$\boldsymbol{\mu}_i(t + \Delta t) = \boldsymbol{\mu}_i(t) + \mathbf{v}_i \Delta t, \quad \Sigma_i(t + \Delta t) = R_i(\Delta t) \Sigma_i(t) R_i(\Delta t)^\top,$$

so prediction is a transformation of the same primitives rather than the synthesis of a new image. This is what "grounded" means here: the prediction is expressed in the geometry, and the predicted state is again an inspectable field.

Whether this composition amounts to a world model, and not merely an interpolation of observed motion, is the substantive question, and recent systems in autonomous driving take it seriously. GaussianWorld reformulates 3D occupancy prediction as forecasting the next state

of an explicit Gaussian scene, decomposing scene evolution into ego motion, the local movement of dynamic objects, and the completion of newly observed regions, and reports an improvement over its single-frame counterpart from the world-model formulation^[6]. It is, to the definition of this report, a spatial world model in the intended sense: an explicit field that predicts. A parallel and larger line pursues generalizable ego-vision world models that generate controllable future observations conditioned on ego trajectory and object dynamics^[7]; these are powerful, but their state is a generated image sequence rather than an explicit field, and by the criterion of Section 3 they satisfy the dynamical half of the definition without the representational half. The distinction is not a ranking. It marks where a system falls on the two axes, and Spatial AI as defined here is the corner where both are satisfied.

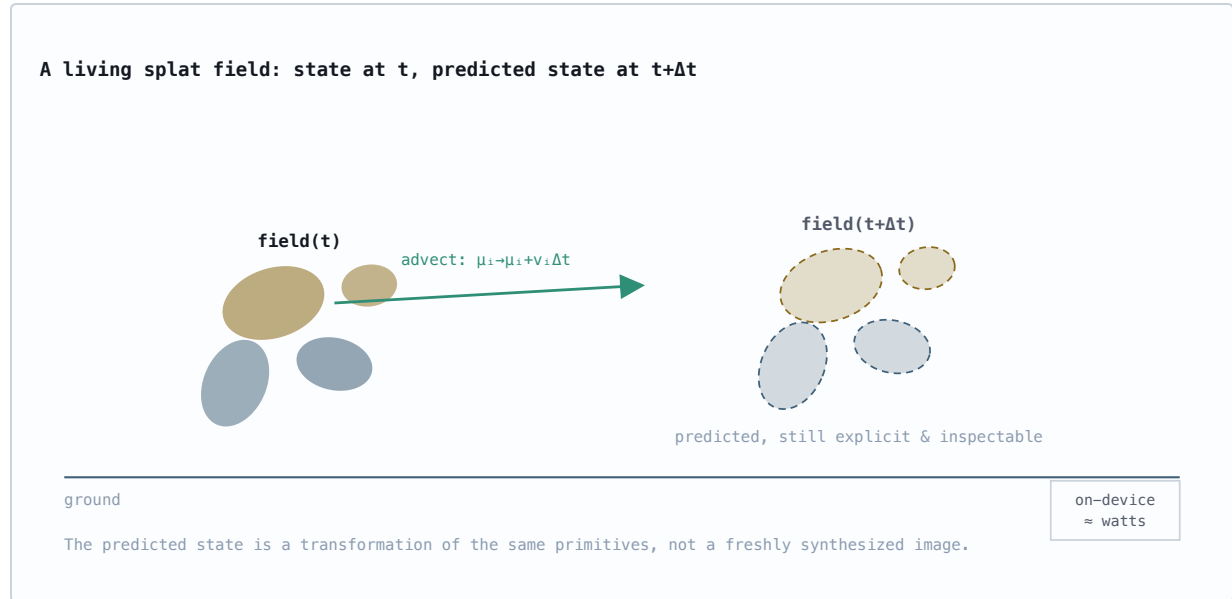


Figure 1. The two halves of the definition in one picture. The explicit Gaussian field at time t (solid) is advected by grounded per-primitive dynamics into a predicted field at $t+\Delta t$ (dashed ghost), which remains an explicit, inspectable field rather than a rendered frame. The Swap-2C constraint requires the whole loop to run on the embodied device at a watt-scale power budget. Illustrative schematic, not measured data. An interactive version is at physicalai-bmi.org/research/charlot-lab.

5. The edge constraint: a world model the size of a robot

The systems in Section 4 largely assume the datacenter. Fitting a splat field, training a deformation model, and running a generative world model are, in their published forms, offline or server-side computations. The lab's constraint reverses this. A robot models the room it is standing in, on the processor it carries, while it acts. The world model must therefore be the size of the robot, not the size of a server farm. The lab labels this constraint Swap-2C: the model runs on the low-power compute that lives in the world it is modeling, and it swaps against a fixed on-board power and memory budget rather than against a datacenter's supply.

This is a hard constraint but not an idle one, because on-device splatting is exactly where the representation's explicitness pays off. Rasterizing an explicit field is far cheaper than marching rays through an implicit one, which is why real-time rendering was the headline result of the

original method^[2]. Recent work on on-device Gaussian splatting reports pipelines targeting mobile and embedded hardware, trading primitive count and precision for a fit within a phone-class power and memory envelope^[8]. The lab treats that result as the existence proof for the static half of the loop on the edge; the open work is to carry the dynamic half — prediction, not just rendering — into the same budget. The levers are the ones the explicit representation exposes: prune primitives aggressively where the scene is simple, quantize the per-primitive parameters, restrict the dynamics to the small set of Gaussians that actually move, and predict only over the horizon control needs. None of these is available to a monolithic latent model, because it has no primitives to prune or freeze. The claim of this section is a design position rather than a benchmarked system: the same property that makes the field inspectable, its explicitness, is what makes an edge budget reachable.

6. Accounting in joules: the MathGround substrate

An edge constraint is only meaningful if it is measured, and the correct unit is not frames per second but joules per predicted field. The energy cost of computation at the edge is dominated not by arithmetic alone but by the movement of operands between memory and the arithmetic units, a cost that at scale exceeds the arithmetic itself^[9]. A splat field is a data-movement problem before it is an arithmetic one: every rendered frame and every predicted step reads and writes the primitives. This is why the lab accounts the whole loop in joules on its MathGround substrate, which prices each operation and each memory transaction rather than counting only multiply-accumulates. The intent is that the fit, the render, and the prediction are all denominated in the same energy currency, so a design change that halves the primitive count or freezes the static Gaussians shows up directly as joules saved rather than as an abstract reduction in floating-point operations.

The unit of a spatial world model at the edge is joules per predicted field. A representation earns its place not only by fidelity but by what it costs to render and to advance one step in time, on the device, within the body's power budget.

The energy account also disciplines the dynamics. A generative world model that redraws the entire scene each step pays for every pixel it hallucinates. A grounded field that advects the primitives that moved pays only for the motion. On an explicit representation the second is expressible; on a latent one it is not. MathGround is, at present, an accounting and simulation substrate within the lab rather than a fabricated device, and the numbers it will produce are design targets, not measured results. Its role in this report is to fix the metric by which the programme judges itself.

7. Position and discussion

The position of this report is a definition and a constraint. Spatial AI, to be a research object rather than a slogan, is the composition of an explicit field and grounded dynamics: a living

splat field that predicts, held in a representation one can inspect, edit, and measure. Table 1 places the surrounding systems on the two axes the definition names.

Table 1. Where representative systems fall on the two axes of the definition.

SYSTEM / CLASS	EXPLICIT FIELD?	GROUNDING DYNAMICS?	EDGE / ON-DEVICE?
NeRF ^[3]	No (implicit weights)	No (static)	Costly (ray marching)
3D Gaussian splatting ^[2]	Yes	No (static scan)	Real-time render
Dynamic / 4D splatting ^[4,5]	Yes	Partial (observed motion)	Mostly offline fit
GaussianWorld ^[6]	Yes (Gaussians)	Yes (forecasts scene)	Server-side
Ego-vision world model ^[7]	No (generated frames)	Yes (controllable)	Datacenter
On-device splatting ^[8]	Yes	No (static, so far)	Yes (phone-class)
Spatial AI (this report's target)	Yes	Yes	Yes (Swap-2C)

Read across the table, the field has already produced explicit fields, grounded prediction, and on-device rendering, but not yet in one system. Commercial efforts to build large persistent 3D world models, such as those pursued by World Labs^[10], sit in the top rows: rich explicit worlds generated and served from the datacenter. The corner this report defines, all three properties at once on the embodied device, is unoccupied, and that is the lab's programme. The argument that it is reachable rests on a single structural fact rather than on optimism: the explicitness that makes the field a world model one can inspect is the same explicitness that makes it prunable, freezable, and cheap to advance one step, which is what an edge budget requires. Whether the fidelity survives that compression is the open empirical question, and this report does not claim to have answered it.

Two caveats bound the position. First, "grounded dynamics" is a spectrum, from pure advection of observed motion to genuine forecasting of unobserved futures, and the harder end is not solved on any representation, least of all at the edge. Second, the explicit field is not free of failure modes: it can overfit to captured views, and it does not by itself supply semantics. The definition offered here is a frame for the work, not a claim that the work is done.

8. Conclusion

Spatial AI has been a buzzword because it named an ambition without naming a mechanism. This report supplies the mechanism. A spatial model holds the world as an explicit field of Gaussian splats, which keeps the geometry and appearance real and the state inspectable, and it couples that field to grounded dynamics, which turn a scan into a world model that predicts. The composition is a living splat field. The constraint that makes it a lab programme rather

than a survey is that it must run at the edge, on the device that lives in the scene, and be accounted honestly in joules on the MathGround substrate. The near-term work is not to invent a new representation but to carry an existing explicit one, and its grounded dynamics, into the power budget of a body. A world model the size of a robot, not a server farm, is the target, and the explicitness of the field is the reason it is within reach.

References

1. D. Ha, J. Schmidhuber. World Models. arXiv:1803.10122, 2018 (NeurIPS 2018).
2. B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graphics 42(4), SIGGRAPH 2023. doi:10.1145/3592433.
3. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:2003.08934, 2020 (ECCV 2020).
4. J. Luiten, G. Kopanas, B. Leibe, D. Ramanan. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. arXiv:2308.09713, 2023.
5. G. Wu, T. Yi, J. Fang, et al. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. arXiv:2310.08528, 2023 (CVPR 2024).
6. S. Zuo, W. Zheng, Y. Huang, J. Zhou, J. Lu. GaussianWorld: Gaussian World Model for Streaming 3D Occupancy Prediction. arXiv:2412.10373, 2024 (CVPR 2025).
7. M. Hassan, S. Stapf, A. Rahimi, et al. GEM: A Generalizable Ego-Vision Multimodal World Model for Fine-Grained Ego-Motion, Object Dynamics, and Scene Composition Control. arXiv:2412.11198, 2024 (CVPR 2025).
8. Mobile-GS: On-Device Gaussian Splatting for Embedded and Mobile Rendering. arXiv:2603.11531, 2026.
9. M. Horowitz. Computing's energy problem (and what we can do about it). IEEE ISSCC, 2014. doi:10.1109/ISSCC.2014.6757323.
10. World Labs. Large World Models. worldlabs.ai (company technical materials, accessed 2026).

AI-use disclosure. Preparation of this report used a large language model (Claude, Anthropic) for drafting and editing text, organizing the reviewed literature, and preparing the figures and the interactive companion. Cited references were checked to resolve to their sources. The author reviewed the content and is solely responsible for it. Consistent with ICMJE, COPE, and IEEE guidance, the model is a tool and is not credited as an author.