Geometry-Centered 3D Latent World Models for Growing Surfaces

Xiaoyi Liu
Department of Computer Science
Brown University
xiaoyi_liu@brown.edu

Hao Tang[†]
School of Computer Science
Peking University
haotang@pku.edu.cn

Abstract

Many physical systems do not merely move or deform; they grow, adding material and changing the geometry that a world model must represent. Existing world models are typically optimized for pixel prediction, reward prediction, or fixed-support physical dynamics, leaving open how to model systems whose underlying physical support expands over time and whose future morphology depends on hidden material response. We introduce FOLIAGE, a geometry-centered latent world model for growing surfaces. Within a fixed state budget, FOLIAGE represents mature regions as a compact scaffold while allocating higher-resolution state to regions predicted to drive near-future growth. This focuses representation and computation where new material and geometric change occur while retaining compact global context. FOLIAGE further separates observation, action, and privileged physics: heterogeneous RGB, point-cloud, and mesh observations are fused into a deployable geometric state; material controls condition the latent dynamics; and hidden physical energies guide training but are not required at deployment. To evaluate this setting, we introduce SURF-GARDEN and SURF-BENCH, providing controlled counterfactual branches, dense cross-modal correspondences, hidden physical signals, and stress tests for growing-geometry state learning. FOLIAGE reduces inverse-material error by $\approx 40\%$ and 5-step mesh forecasting Chamfer error by $\approx 30\%$ relative to strong baselines, while improving cross-modal retrieval by +14 mAP points. Stress tests show graceful degradation under sensor loss and correspondence corruption. On temporal 3D plant scans, FOLIAGE also improves passive future-geometry forecasting, while transfer experiments show that the learned geometry-centered state remains useful beyond the simulator.

1 Introduction

Many physical systems take shape by growing. Leaves curl, tissues fold, and active or printed sheets change form as local material response varies across a surface. In such systems, geometry is not merely displaced; the surface itself expands, new degrees of freedom appear, and local growth can propagate into global morphology. Modeling this process matters because *growth is a mechanism for creating form and function*: it determines how biological structures develop, how soft materials bend and fold, and how programmable surfaces can be designed to reach desired shapes in fabrication. Liang and Mahadevan [2011], Coen and Cosgrove [2023], van Rees et al. [2017a], Gladman et al. [2016], Wang et al. [2024], van Manen et al. [2021], Ambrosi et al. [2019]. It also poses a fundamental machine-learning question: how should a learned world model represent a physical state whose geometry grows over time? We study this question in thin growing surfaces, where local changes in surface material can be observed as large-scale changes in shape. This setting is simple enough to

[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

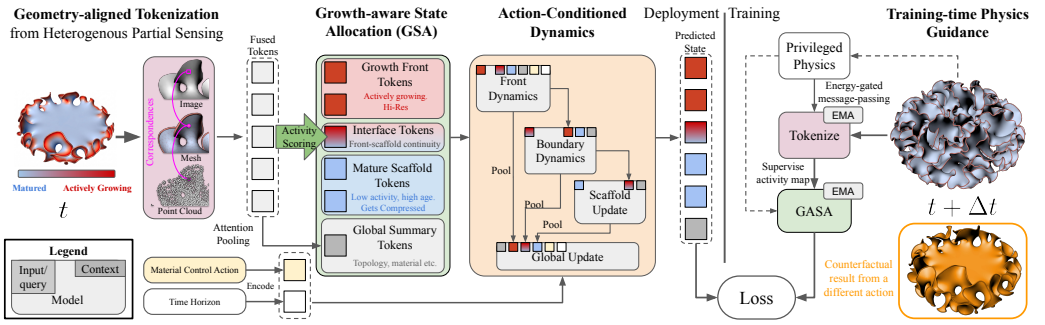


Figure 1: FOLIAGE models growing surfaces by building a geometric state from partial sensors, allocating a fixed state budget according to estimated growth activity, and rolling the state forward under material-control actions. Privileged physical signals guide training only (not required at deployment).

define geometrically, yet rich enough to require tracking where new surface appears, how older and newer regions interact, and how hidden material response shapes future geometry.

Related approaches partially address this problem, but begin from different state assumptions. Video and latent world models learn how observations evolve over time, so they can capture visible evidence of growth Hafner et al. [2019, 2023]. For growing surfaces, however, visible change is only a partial view of physical change: internal stress that shapes the surface is hidden, and an active region may curl, self-occlude, or leave the current view while continuing to affect future geometry. Physics-based simulators and learned physical dynamics model the evolution of physical states represented by meshes, particles, or fields Du et al. [2021], Li et al. [2022]. These approaches are powerful when the current physical state and discretization are given. Growing surfaces add a complementary state-estimation problem: the current geometry may be only partially observed, while future material support, birth locations, and connectivity are not known in advance. Static 3D representation and correspondence methods track geometric identity, but do not model how growth changes under material interventions Melzi et al. [2019], Sharp et al. [2022]. These perspectives suggest the state abstraction needed for growing geometry: *a geometric representation of the evolving surface that can be inferred from heterogeneous partial observations, updated under material controls, and shaped during training by hidden physical signals when available.*

We introduce FOLIAGE, a geometry-centered latent world model for growing surfaces (Fig. 1). Its central algorithmic idea is *growth-adaptive state allocation* coupled with structure-matched dynamics. The physical surface represented by the model can gain material and geometric degrees of freedom over time, but representing this growth does not require the latent state to grow with the mesh. Instead, FOLIAGE uses a fixed state budget and assigns that capacity according to where future geometric change is expected. Mature regions are represented as a compact scaffold that preserves global shape, topology, boundary conditions, and long-range constraints. Young or high-activity regions receive denser growth-front tokens, while interface tokens preserve continuity between the active front and the mature surface. The dynamics model then updates these roles differently, spending its most expressive computation on future-active regions while retaining the surrounding scaffold as context. This gives a practical inductive bias for growing surfaces: as the underlying geometry expands, state capacity should follow the parts of that geometry that matter most for its future evolution.

The next challenge is to preserve the state representation's counterfactual meaning: from the same observed surface, the model should predict different futures under different material controls. FOLIAGE therefore keeps three roles separate. Observations from RGB images, point clouds, and meshes infer the current growing-surface state. Material controls enter only through the latent dynamics, so they intervene on the future rather than changing the meaning of the current state. Hidden physical quantities, such as per-vertex energies that summarize internal stress, guide representation learning during training but are excluded from the deployed path. This role separation lets FOLIAGE use privileged physics to learn better states while preserving a practical observation-based model for action-conditioned geometric prediction. In particular, one inferred current state can be queried under multiple material controls, so the action selects a possible future rather than changing the representation of the observed present.

To evaluate growing-geometry world models, visual prediction alone is insufficient. A useful state should preserve global geometry, track identity through growth, align heterogeneous sensors, infer

hidden material causes, predict counterfactual futures, and remain usable under partial or noisy observations. We therefore introduce SURF-GARDEN, a controlled data platform for generating growing surface sequences, and SURF-BENCH, an evaluation suite organized around these requirements. SURF-GARDEN provides counterfactual branches, dense temporal and cross-modal correspondences, and hidden physical signals that are difficult to measure at scale in real growth data. The task suite probes whether the latent behaves like a physical state of a growing object: topology and dense correspondence test geometry understanding; image–point retrieval and sensor ablations test multimodal grounding; inverse material estimation and growth-stage recognition test hidden physical fidelity; and mesh forecasting and long-horizon rollout test action-conditioned dynamics. Across these tasks and stress tests, FOLIAGE outperforms task-specific baselines while degrading gracefully under sensor and correspondence corruption. Real plant experiments additionally execute the learned transition model for passive future-geometry prediction, while plant and dynamic-human transfer experiments test whether the learned representation remains useful beyond SURF-GARDEN.

In summary, our contributions are:

- **Growing geometry as a world-modeling problem.** We formulate growing surfaces as a regime where the underlying physical support changes over time, current geometry must be inferred from partial observations, material controls alter future morphology, and future support is not known in advance.
- **Growth-adaptive geometric state and structured dynamics.** We introduce FOLIAGE's central algorithmic mechanism: a fixed state budget is organized into mature scaffold, active growth-front, and interface tokens, with role-specific dynamics that concentrate representation and computation on predicted future-active regions.
- **Counterfactual interface with train-only physics.** FOLIAGE forms the current state from observations, introduces material interventions in the transition, and confines privileged physical signals to training. This allows one observed state to support multiple action-conditioned futures without requiring solver quantities at deployment.
- **Controlled benchmark for growing-geometry state learning.** We introduce SURF-GARDEN and SURF-BENCH, providing counterfactual branches, dense correspondences, hidden physical signals, and stress tests for evaluating geometry, identity, multimodal grounding, physical fidelity, and predictive dynamics.
- **Empirical evidence across geometry, dynamics, robustness, and real-world forecasting.** FOLIAGE improves geometric understanding, multimodal grounding, hidden-material inference, and forecasting; matched-budget controls isolate the benefit of growth-aware allocation and structured dynamics; and temporal plant scans demonstrate passive future-geometry prediction beyond the simulator.

2 Related Works

Video and latent world models. Latent world models learn compact states for prediction, planning, and control from high-dimensional observations [Ha and Schmidhuber, 2018, Hafner et al., 2019, 2023]. Most work focuses on image-centric settings, where the state is trained via reconstruction or policy objectives. In parallel, neural scene representations and 3D perception emphasize explicit spatial structure for view synthesis, reconstruction, and embodied reasoning [Mildenhall et al., 2021, Kerbl et al., 2023, Jaegle et al., 2021]. FOLIAGE follows this geometry-centered direction but targets a distinct regime: geometry is not just a representation, but the evolving physical state. Growing surfaces require tracking material identity, support, and future morphology under control, making geometry itself the object of prediction.

Learned physical dynamics and growing support. Differentiable simulators and neural physical models enable inverse problems, control, and forecasting when states and assumptions are specified [Hu et al., 2020, 2019, Du et al., 2021, Li et al., 2022]. Graph-based simulators and operator-learning methods further model dynamics over particles, meshes, and fields with strong generalization [Sanchez-Gonzalez et al., 2020, Pfaff et al., 2020, Li et al., 2024b]. However, growing surfaces introduce changing state support and partially observed internal physics. FOLIAGE addresses this by learning latent dynamics over geometry, using simulator signals only during training. This preserves physical structure without requiring solver access at deployment.

Multimodal geometry, correspondence, and physical identity. Multimodal 3D learning aligns images, point clouds, and meshes for retrieval and recognition [Afham et al., 2022, Zhang et al., 2022, Xue et al., 2024], while correspondence and functional-map methods match surfaces across shapes [Ovsjanikov et al., 2012, Melzi et al., 2019, Donati et al., 2020, Sharp et al., 2022]. In growing geometry, correspondence extends beyond static matching to track physical identity as surfaces evolve. FOLIAGE uses correspondence-aware fusion to build a shared geometric state from partial observations, ensuring it supports both static tasks and future prediction under growth.

Simulation, counterfactual data, and privileged training signals. Simulation provides scalable supervision, controlled interventions, and privileged variables [Tobin et al., 2017, James et al., 2019, Makoviychuk et al., 2021, Szot et al., 2021]. This is especially valuable for growing geometry, where counterfactual futures, dense correspondences, and internal energies are difficult to observe. Related ideas in privileged learning and distillation leverage additional training signals to improve deployment representations [Vapnik and Izmailov, 2015, Hinton et al., 2015, Tarvainen and Valpola, 2017]. FOLIAGE follows this paradigm: per-vertex energies guide training, while deployment relies only on observable sensors. SURF-GARDEN and SURF-BENCH provide controlled evaluation with counterfactual growth and robustness tests.

Growing surfaces and morphology. Growth-driven shape formation has been widely studied in mechanics, morphogenesis, and active materials [Efrati et al., 2009, Liang and Mahadevan, 2011, Gladman et al., 2016, van Rees et al., 2017b, Ambrosi et al., 2019]. These works show how local growth and material response produce complex global morphology. FOLIAGE brings this regime into learned world modeling, focusing on inferring compact states from partial observations and predicting their evolution under control. More broadly, when systems gain degrees of freedom over time, effective world models benefit from growth-adaptive state allocation that separates stable structure from regions of active change.

3 FOLIAGE: A Growth-Adaptive World Model for Growing Surfaces

3.1 Problem formulation and interfaces

FOLIAGE uses a world-model interface specialized to growing surface geometry. At time t, the model receives any *subset* of RGB images, point clouds, and surface meshes, $x_t \subseteq \{I_t, P_t, M_t\}$, and infers a current geometric state S_t. Given a material action $a_t \in \mathbb{R}^3$ and prediction horizon Δt, its transition model predicts $\hat{S}_{t+\Delta t} = P_\theta(S_t, a_t, \Delta t)$. When no intervention is supplied, a learned null-action token supports passive prediction. The growing-surface setting changes what this interface must represent: the future surface may contain material and geometric degrees of freedom that do not exist at time t, and its future support, birth locations, and connectivity are not provided to the model.

Importantly, changing physical support does not require the number of latent tokens to increase with the surface. FOLIAGE instead uses fixed scaffold, front, and interface budgets. At each encoded observation, growth-adaptive state allocation determines which parts of the observed surface receive detailed state. The transition then predicts a future structured state supervised by a target state formed from the future, expanded surface. Thus, the amount of represented physical material may increase substantially from t to $t + \Delta t$, while the latent representation remains computationally controlled and changes what geometric information its capacity represents.

Abstraction. FOLIAGE represents the current surface by the structured state $S_t = \{s_t^{\mathrm{global}}, S_t^{\mathrm{mat}}, S_t^{\mathrm{front}}, S_t^{\mathrm{bdry}}\}$. s_t^{global} summarizes the whole surface, S_t^{mat} stores compact tokens for mature scaffold geometry, S_t^{front} stores higher-resolution tokens for regions predicted to drive near-future change, and S_t^{bdry} stores interface tokens connecting the active front to the scaffold. This decomposition is the central interface of FOLIAGE: growing geometry requires stable global context together with expressive local state where new material and deformation are likely to appear.

The method is organized around four subproblems. First, heterogeneous observations are converted into geometry-aligned tokens (Sec. 3.2). Second, those tokens are allocated into mature scaffold, active front, and interface state (Sec. 3.3). Third, the structured state is updated under material controls while keeping current-state inference observational (Sec. 3.4). Fourth, hidden physical quantities improve training without becoming test-time requirements (Sec. 3.5).

3.2 Build a geometric state from heterogeneous partial observations

A growing-geometry world model needs a state whose coordinates are geometric rather than purely visual. Since observations may arrive as RGB images, point clouds, meshes, or subsets thereof, FOLIAGE first maps each available modality into a shared token space $U_t = T_I \cup T_P \cup T_M$. Missing modalities contribute empty token sets, so downstream modules operate on the same representation regardless of which sensors are present. Image tokens provide appearance cues, point tokens provide metric anchors, and mesh tokens provide surface structure.

The part of tokenization most specific to growing geometry is the mesh pathway. Mesh vertices are not all equivalent: some correspond to older material that mainly anchors the global scaffold, while others were recently added and are more likely to indicate active growth. Each mesh token therefore carries geometric features and a birth-time tag $\tau(v)$, from which we derive an age feature $\alpha_t(v) = t - \tau(v)$. Image and point tokens inherit age information through nearby mesh correspondences when available. This lets the model reason about the age structure of the surface even when the current observation is an image or point cloud.

FOLIAGE fuses modalities through a sparse correspondence graph rather than dense all-to-all attention. The graph contains image-to-mesh edges from projected visible surface elements, point-to-mesh edges from local geometric proximity, and mesh adjacency edges from surface connectivity. Type-aware attention over this graph lets image appearance refine geometric tokens and point measurements supply metric evidence, while keeping all modalities tied to the same physical surface. Because practical sensing can be partial or misaligned, training uses token masking, paired masking along correspondence edges, and modality dropout. Thus, correspondences are used as a useful geometric interface, not as a deployment assumption that every observation is complete or perfectly registered. We test this in Sec. 4.4

3.3 Allocate state capacity according to growth

Growing surfaces are not uniformly dynamic. Mature regions often preserve global shape, topology, and boundary context, while material addition and control-sensitive deformation concentrate near active growth fronts. A single pooled latent risks hiding these differences: it may underrepresent the active regions where the future changes, or spend unnecessary capacity on mature regions whose near-term dynamics are small. FOLIAGE therefore performs **Growth-adaptive State Allocation (GSA)**: it compresses mature geometry into scaffold tokens, retains high-resolution tokens for active growth fronts, and adds interface tokens that keep local growth attached to global structure. Here, "higher resolution" refers to allocating a denser share of the fixed latent-state budget, together with a more expressive dynamics update, to predicted future-active regions. It does not refer to increasing the resolution of the output mesh itself.

Given fused tokens U_t, FOLIAGE first predicts an observable activity score $g_i = \sigma\big(f_\psi(u_i)\big)$ for each token. The score estimates where near-future change is likely. It is computed from observable token features, including geometry, age, modality, and correspondence context. During training, hidden physical quantities and future local change can supervise this score, but they are not inputs to the deployed selector. This preserves the main interface: FOLIAGE can learn from privileged physics while still forming its state from observable sensors alone.

The activity scores define three complementary token sets for S_t. Low-activity or older regions are coarsened into *mature scaffold tokens* S_t^{mat}, which preserve global geometry, topology, and long-range constraints. High-activity regions become *growth-front tokens* S_t^{front}, which receive the highest-resolution representation because they are where new degrees of freedom are most likely to appear. Tokens near the transition between front and scaffold form *interface tokens* S_t^{bdry}, which preserve attachment between local growth and the surrounding surface. Finally, attention pooling over all fused tokens produces s_t^{global}, a compact summary of overall shape and growth stage.

This decomposition is the main algorithmic mechanism of FOLIAGE. As the physical surface gains new support, the model must decide where a fixed amount of state capacity is most useful. GSA makes this decision from observable evidence: active regions receive detailed state, mature structure remains compact, and interface tokens preserve their attachment. The same structure then determines how dynamics are applied, with the strongest updates at the front and progressively lighter updates toward the mature scaffold. In Sec. 4.3, we compare against random and incomplete decompositions

as well as same-budget uniform sampling, learned latent queries, generic learned pruning, and growth-aware allocation with homogeneous dynamics. These controls separate the effect of growth-specific allocation from additional capacity or generic adaptivity.

3.4 Predict future geometry under material control actions

A useful model for growing surfaces should support counterfactual prediction: the same observed state can evolve differently under different material controls. FOLIAGE therefore separates **observation** from **action**. The perception and allocation modules form S_t using only observable inputs; material controls enter only in the dynamics model: $\hat{S}_{t+\Delta t} = P_\theta(S_t, a_t, \Delta t)$. This separation preserves the state's meaning: S_t describes the current growing surface, and a_t selects a possible future.

The dynamics model respects the structure introduced by GSA. Active growth-front tokens receive the most expressive action-conditioned update, because they represent regions where new material and near-future deformation are concentrated. Interface tokens propagate front changes into the surrounding surface, while mature scaffold tokens receive a lightweight residual update that preserves global geometry and long-range context. This gives the model a simple inductive bias: actions should influence the parts of the surface most likely to change, but those changes should remain attached to the mature scaffold. Material action a_t is encoded as an action token, and the horizon Δt a time token. When material control actions are unavailable, a learned null-action token supports passive prediction. Multi-horizon training exposes the model to short and longer prediction intervals, allowing the same transition model to support direct counterfactual prediction and recurrent rollouts.

3.5 Use hidden physics as train-time guidance

The future of a growing surface depends on internal stress and material response, but these quantities are generally not available at deployment. FOLIAGE therefore separates **privileged physics** from both observation and action: hidden physical signals guide representation learning during training, while the deployed model uses observable inputs alone.

We maintain two encoders. The **context encoder** consumes only observable inputs (Sec. 3.2) and produces the deployable state S_t. The **target encoder** is used only during training; it receives the same observations plus simulator-provided per-vertex physical energies and produces the target future state used to supervise prediction. The target encoder is updated as an exponential moving average of the context encoder, and prediction gradients do not flow through privileged physical inputs.

Physical energies are used in two ways. First, in the target encoder, they gate early mesh message passing so that high-stress regions contribute more strongly to the target representation. Second, they supervise the activity map used by GSA, together with future local change, teaching the observable context branch where growth-driven change is likely to occur. These signals improve the learned state without changing the test-time interface: at inference, FOLIAGE receives only x_t, forms S_t, and predicts future states under material actions.

We use several safeguards to keep the training/deployment separation explicit. Energies are detached constants, gates are bounded, physics-guided message passing is applied only in the target branch, and the target branch is EMA-coupled to the observable context branch. Thus, train-only physics shapes the representation, but the deployed model remains solver-free.

3.6 Training and Deployment

FOLIAGE is trained to predict the future structured state produced by the target encoder. The objective contains latent prediction, activity supervision, energy prediction, and representation regularization: $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{act}}\mathcal{L}_{\text{act}} + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}} + \lambda_{\text{vc}}\mathcal{L}_{\text{vc}}$. $\mathcal{L}_{\text{pred}}$ matches the predicted future state to the target state, $\mathcal{L}_{\text{act}}$ teaches the activity selector where future change occurs, $\mathcal{L}_{\text{energy}}$ drives observable features to correlate with hidden stress, and $\mathcal{L}_{\text{vc}}$ prevents representation collapse. The inference path is simpler: observable inputs are tokenized, GSA forms the structured state S_t, and the dynamics model predicts $\hat{S}_{t+\Delta t}$ under material action a_t and horizon Δt. No target encoder, physical energy, or simulator state is used. Downstream task heads read from the same structured state.

For explicit geometry forecasting, a variable-cardinality readout converts $\hat{S}_{t+\Delta t}$ back to surface geometry. Queries anchored to the current surface predict updates to existing geometry, while learned

Table 1: FOLIAGE preserves global and local geometric identity and aligns heterogeneous sensing modalities. For retrieval, we report image (I)→points (P), zero-shot I→P, and zero-shot P→I.

Topology Classification		**Dense Correspondence**		**Cross-Modal Retrieval**	
	%Accuracy↑		Geodesic Err.↓		mAP@100 (%)↑
MeshCNN Hanocka et al. [2019]	88	ZoomOut Melzi et al. [2019]	4.2	CrossPoint Afham et al. [2022]	42/18/16
DiffusionNet Sharp et al. [2022]	92	DiffusionNet Sharp et al. [2022]	3.8	CLIP2Point Huang et al. [2023]	43/20/19
Nishikawa et al. [2023]	94	G-MSM Eisenberger et al. [2023]	3.6	ULIP-2 Xue et al. [2024]	46/23/21
ETNN Battiloro et al. [2024]	94	Cao et al. [2024a]	3.2	PointCLIPv2 Zhang et al. [2022]	48/22/23
Ours	**98**	**Ours**	**2.6**	**Ours**	**64/39/37**



Source Target (GT) Ours Cao et al. [2024a] DiffusionNet G-MSM ZoomOut

Figure 2: FOLIAGE accurately tracks features in dense correspondence tasks as baselines degrade under significant surface morphology changes caused by accretive growth.

growth queries predict candidate new vertices and validity scores; valid additions are reconstructed locally. Future vertex count, birth locations, and connectivity are therefore not provided as inputs. This readout tests whether the predicted state contains the information needed to reconstruct surface that did not exist at the current time.

The structured dynamics themselves use fixed scaffold, front, and interface token budgets rather than scaling with full mesh resolution. Growth is therefore represented through the geometric content and organization of the latent state, together with the variable-cardinality geometry that state supports, rather than by increasing the cardinality of the recurrent latent tensor.

4 Results

We evaluate whether FOLIAGE learns a deployable geometric state for growing surfaces. Such a state should preserve global geometry, track identity through growth, ground heterogeneous sensors, encode hidden material causes, and support future prediction under material actions. SURF-GARDEN provides the controlled setting needed to test these properties through matched counterfactual branches, dense temporal and cross-modal correspondences, hidden physical signals, and complete surface geometry. SURF-BENCH then probes the learned state through capability-focused tasks: geometry and multimodal grounding in Sec. 4.1, hidden-cause modeling and action-conditioned forecasting in Sec. 4.2, architectural ablations in Sec. 4.3, and robustness and real-world generalization in Sec. 4.4.

4.1 Does FOLIAGE preserve geometry and identity for growing surfaces?

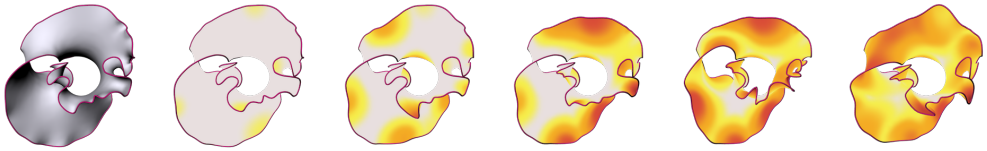
If FOLIAGE learns geometry as state, it should preserve both global surface structure and local identity as new material appears. We test this in two ways: topology classification tests whether the latent retains global geometric invariants; dense correspondence tests whether it tracks the same physical or semantic regions through growth. These are state tests: a model that predicts plausible local shape but loses topology or correspondence has not learned a reliable physical state.

In Tab. 1, FOLIAGE significantly improves % topology accuracy from 94 to 98 and reduces dense-correspondence geodesic error from 3.2 to 2.6. Fig. 2 gives the qualitative counterpart: FOLIAGE tracks features through growth-induced morphology changes more consistently than static correspondence baselines. Together, these results show that the learned state preserves both what the surface is globally and how its regions persist locally through growth.

Multimodal grounding under partial sensors. If FOLIAGE builds one geometric state from partial sensors, RGB and 3D observations of the same growing surface should align in the learned space. We test this with image–point retrieval: the question is whether heterogeneous observations are grounded to the same physical surface state. In Tab. 1, FOLIAGE substantially improves supervised image–point retrieval; in zero-shot retrieval, FOLIAGE also remains ahead in both image-to-point and point-to-image directions. These gains indicate that correspondence-aware geometric fusion

Table 2: FOLIAGE captures hidden causes and models counterfactual futures

Inverse Material Regression MAE($\times 10^3$) ↓		**Growth-Stage Recognition** %Balanced Acc.↑		**Mesh Forecasting (direct 5-step)** Chamfer($\times 10^3$) ↓ / Vertex Drift↓	
Kairanda et al. [2024]	60	SVFormer Xing et al. [2023]	67	MeshGraphNets Pfaff et al. [2020]	65/4133
DiffPD Du et al. [2021]	58	VideoMAE-v2 Wang et al. [2023]	68	CaDeX Lei and Daniilidis [2022]	52/2261
BDP Gong et al. [2024]	55	TimeSformer Bertasius et al. [2021]	69	MeshGPT-solo Siddiqui et al. [2024]	45/2540
DiffCloth Li et al. [2022]	53	VideoMamba Li et al. [2024a]	71	Motion2VecSets Cao et al. [2024b]	38/1721
Ours	**32**	**Ours**	**82**	**Ours**	**26/915**



Ground Truth Ours Motion2VecSets MeshGPT-solo CaDeX MeshGraphNets

Figure 3: Mesh predictions over 10 time steps under control actions. Higher errors = more red. FOLIAGE's predictions closely resemble the ground truth while baselines' rapidly deviate.

produces a shared state across modalities, rather than separate image and 3D embeddings that only work under task-specific supervision.

4.2 Is FOLIAGE a good world model? Hidden causes and counterfactual futures

If FOLIAGE learns a physical state for growing geometry, the state should encode hidden causes of future morphology and support different futures under different material actions. We test this in three ways: material estimation probes whether the latent captures physical factors that drive growth; stage recognition asks whether it captures where the process is in its trajectory; forecasting tests whether the state can be rolled forward under action-conditioned dynamics. In Tab. 2, FOLIAGE reduces inverse-material mean absolute error and 5-step mesh-forecasting Chamfer distance, while also improving growth-stage recognition and vertex drift. Fig. 3 gives the qualitative counterpart: FOLIAGE better preserves global morphology and active-front detail at the queried future horizon. Together, these results show that the learned state contains both information about hidden physical causes and information needed to predict future geometry.

Does the action actually select the future? We additionally evaluate counterfactual sibling branches that share the same observed prefix but evolve under different material controls. Given the queried action, FOLIAGE retrieves the corresponding ground-truth sibling future in 90.4% of cases and obtains a predicted-to-true branch-separation ratio of 0.95 (1 is ideal). Moving the action into the state encoder instead reduces these values to 78.2% and 0.84. This supports the intended interface: the inferred current state describes the observed surface, while the transition selects among alternative action-conditioned futures.

4.3 Why Does FOLIAGE Work?

FOLIAGE's gains should come from design choices matched to growing geometry, not from undifferentiated model capacity. In Tab. 3, we test this by ablating growth-adaptive state allocation, then by removing the interfaces that ground sensors, localize growth, and use train-only physical signals.

Effect of GSA. Removing GSA and reverting to a flat pooled state hurts forecasting and material inference. Random front selection underperforms despite using a similar token budget, showing that activity-guided allocation matters. Front-only updates lose topology and retrieval performance, showing that local growth must remain anchored to mature geometry. Removing interface tokens weakens forecast consistency. Full FOLIAGE performs best because growth is local but not isolated: active regions must be updated in the context of the mature surface they remain attached to.

Other design choices. Further ablations show that they fail in interpretable ways. Removing correspondence fusion mainly hurts retrieval; removing age/activity features dents growth-sensitive forecasting; removing train-only energies or only energy-gated message passing mainly limits material inference. Each component affects the capability it was introduced to support.

Matched allocation and dynamics. The direct GSA ablation establishes that growth-aware state allocation helps, but does not by itself distinguish growth-specific allocation from generic adaptive

Table 3: Component ablations and matched allocation controls for FOLIAGE. Left and middle report topology classification accuracy (A), inverse-material MAE $\times 10^3$ (M), direct mesh-forecasting Chamfer $\times 10^3$ (C), and retrieval mAP (%) (m). Right isolates state allocation and dynamics under a matched 448-token, approximately 43M-parameter budget; C5 is direct 5-step Chamfer and C40 is recursive 40-step Chamfer, both $\times 10^3$.

Effect of Growth-adaptive state

Variant	A↑	M↓	C↓	m↑
FOLIAGE	**97.7**	**32**	**25.7**	**64.2**
w/o GSA	95.8	35	28.0	63.0
random front	95.6	36	28.7	62.4
front-only	94.9	34	29.4	60.4
w/o interface	96.4	33	26.6	63.6

Effect of interface and supervision

Variant	A↑	M↓	C↓	m↑
FOLIAGE	**97.7**	**32**	**25.7**	**64.2**
w/o corr.	96.1	39	31.8	50.2
w/o masking	96.1	37	29.8	55.8
w/o age feat.	95.2	40	34.4	62.1
w/o energies	96.3	50	28.8	63.3
w/o nrg. gate	96.4	46	28.3	63.1
mean pool	95.5	38	32.0	57.3

Matched allocation & dynamics

Method	Params	GFLOPs	C5↓	C40↓
Uniform FPS	42.8M	17.67	28.2	62.7
Learned queries	43.0M	17.72	27.8	59.8
Generic pruning	43.1M	17.81	27.1	53.6
GSA + flat dyn.	43.0M	17.75	26.4	47.9
FOLIAGE	43.0M	17.82	**25.7**	**41.2**

Table 4: FOLIAGE is robust under partial sensing and imperfect correspondences. (L) We vary modality availability: Rich (all), Typical (RGB + point clouds, incomplete meshes), Sparse (one modality), and Noisy (sensor noise + random modality dropout) and report the best-performing models. (R) We increase correspondence corruption: Corr-noise (local rewiring + edge drop from the multimodal correspondence graph), Occl. (RGB masking with linked 3D tokens), Drift (camera jitter), and SLAM (pose re-estimation).

Sensor-Subset Robustness

Growth-stage recognition Balanced Acc.↑

Method	Rich	Typical	Sparse	Noisy
VideoMAE-v2	0.74	0.68	–	0.62
PiMAE Chen et al. [2023]	0.76	0.70	0.67	0.63
ULIP-2	0.78	0.72	0.71	0.66
CLIP2Point	0.74	0.65	0.68	0.60
Ours	**0.80**	**0.78**	**0.74**	**0.74**

Correspondence Robustness

Cross-modal retrieval mAP@100↑

Method	Clean	Corr.	Occl.	Drift	SLAM
FOLIAGE	**0.60**	**0.56**	**0.51**	**0.52**	**0.59**
FOLIAGE w/o corr.	0.46	0.45	0.43	0.44	0.46
ULIP-2	0.46	0.49	0.40	0.41	0.42
PointCLIPv2	0.48	0.47	0.39	0.40	0.41
CLIP2Point	0.43	0.45	0.37	0.38	0.39
CrossPoint	0.42	0.43	0.35	0.36	0.37

token selection. We therefore compare same-budget alternatives with identical encoders and readouts and matched transition computation. The complete scaffold/front/interface-specific dynamics in FOLIAGE outperforms the alternatives. The gains are greater on C40 than on C5, suggesting that the value of structured allocation becomes clearest under recursive prediction, where local allocation and update errors compound over time.

When is growth-adaptive allocation most useful? GSA is designed for spatially nonuniform change. As active coverage becomes global, its advantage over uniform allocation narrows and activity recall falls, with errors concentrating on missed high-change regions (Appendix J). The same ordering persists through the complete 350-frame rollout (Appendix I). These results characterize GSA's intended regime: adaptive allocation helps most when future change is spatially heterogeneous.

4.4 Does FOLIAGE generalize under imperfect sensing and beyond simulation?

Robustness under realistic sensing conditions. Tab. 4 tests missing modalities and corrupted cross-modal correspondences. FOLIAGE leads across rich, typical, sparse, and noisy sensing, and remains strongest under correspondence noise, occlusion, and camera drift. Performance degrades smoothly as sensing weakens, while the correspondence-free ablation remains usable but lower. Thus, correspondence-aware fusion improves geometric grounding without requiring ideal links, while masking and modality dropout reduce dependence on any single sensor.

We then test whether correspondence-aware fusion requires ideal links (Tab. 4). FOLIAGE remains strongest under correspondence noise, occlusion, and camera drift, and recovers when pose refinement is available. The correspondence-free ablation remains usable but lower. This is expected: correspondences improve grounding when approximate associations exist, but they are not an oracle dependency. These robustness tests show that FOLIAGE's state remains usable under realistic sensing stress like missing modalities, noisy links, and correspondence drift.

Real-world and cross-domain representation transfer. If FOLIAGE learns reusable growing-geometry features rather than simulator-specific shortcuts, its representation should remain useful outside SURF-GARDEN. Tab. 5 tests whether features learned in SURF-GARDEN remain useful

Table 5: Representation transfer beyond SURF-GARDEN. FOLIAGE transfers to **real plant datasets** GrowliFlower Kierdorf et al. [2023], Pheno4D Schunck et al. [2021], and Crops3D Zhu et al. [2024], and to **dynamic human-body datasets** D-FAUST Bogo et al. [2017] and CAPE Ma et al. [2020]. "Frozen" keeps the encoder fixed; "Few-shot" trains a linear probe using 10 labeled sequences per class; "Light-ft" additionally fine-tunes the last encoder block.

Method	**GrowliFlower**
	Growth stage recognition Balanced Acc. ↑
FOLIAGE (Frozen)	0.72
FOLIAGE (Few-shot)	0.75
FOLIAGE (Light-ft)	**0.77**
VideoMAE-v2	0.70
TimeSformer	0.68
VideoMamba	0.69
SVFormer	0.66

Method	**Pheno4D**	**Crops3D**	**D-FAUST**	**CAPE**
	Dense correspondence Geodesic. Err. ↓			
FOLIAGE (Frozen)	3.55	**4.45**	3.6	4.4
FOLIAGE (Light-ft)	**3.25**	–	**3.1**	**3.8**
DiffusionNet	3.90	4.80	3.8	4.7
G-MSM	4.60	5.50	4.5	5.4
Cao et al. [2024a]	5.05	5.95	4.9	5.8
ZoomOut	5.70	6.70	5.6	6.6

Table 6: Passive future-geometry forecasting on real temporal plant scans. Chamfer distance is bounding-box normalized and reported as $\times 10^3$; brackets are 95% bootstrap intervals over held-out plant trajectories after averaging three fine-tuning seeds. Persistence copies the current surface, constant velocity extrapolates motion from the preceding scans, and Uniform AR is a matched variable-cardinality autoregressive model. Splits are by physical plant, and all learned models are lightly fine-tuned on the corresponding training plants.

Method	Pheno4D Maize, 2-day↓	Pheno4D Tomato, 2-day↓	Crops3D Cabbage, 2-week↓
Persistence	32.4 [30.4,34.4]	39.8 [37.2,42.4]	50.8 [47.4,54.2]
Constant velocity	27.6 [25.8,29.4]	33.6 [31.3,35.9]	43.2 [40.2,46.2]
Uniform AR, light-ft	23.1 [21.6,24.6]	28.6 [26.6,30.6]	37.2 [34.6,39.8]
FOLIAGE, light-ft	**19.6 [18.3,20.9]**	**24.2 [22.5,25.9]**	**31.3 [29.1,33.5]**

on independently collected data. FOLIAGE improves real-plant growth-stage recognition and transfers to dense correspondence on plant and dynamic human-body surfaces. These experiments test representation transfer rather than the learned transition itself. Hence, for **real-world future-geometry forecasting,** we execute the transition model on temporal 3D plant scans (Tab. 6). With no intervention labels, FOLIAGE uses its passive null-action token and improves future-geometry prediction across all three settings, including over a matched Uniform AR baseline.

5 Scope and Limitations

FOLIAGE targets *accretive thin surfaces*, rather than arbitrary volumetric growth or unrestricted topology changes such as tearing and merging. Within this regime, we vary topology, material response, sensing, growth localization, rollout horizon, and controls, and additionally evaluate passive future prediction on real plant scans. Intervention-conditioned counterfactuals are evaluated in simulation because available real datasets do not provide matched futures under controlled material actions, leaving real closed-loop control outside the demonstrated scope. Performance also degrades under globally distributed activity, sufficiently poor sensing, and long open-loop rollouts, as quantified by our stress tests. Richer real 3D capture, online state refresh, real intervention datasets, and extensions to more general growth processes are natural directions beyond the setting studied here.

6 Conclusion

We introduced FOLIAGE, a geometry-centered latent world model for growing surfaces. FOLIAGE organizes a fixed state budget into mature scaffold, future-active front, and interface tokens, then applies role-specific dynamics while using privileged physics only during training. Across SURF-BENCH, matched allocation controls, robustness tests, and real plant forecasting, this growth-adaptive state improves geometric representation and future prediction. Together, the controlled counterfactual experiments and real passive forecasting show that the learned state remains useful both for reasoning about alternative growth futures and for predicting geometry beyond the simulator. These results support a simple principle for growing systems: representation capacity should follow where new physical structure is emerging while preserving compact global context.

References

Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9902–9912, 2022.

Davide Ambrosi, Martine Ben Amar, Christian J. Cyron, Antonio DeSimone, Alain Goriely, Jay D. Humphrey, and Ellen Kuhl. Growth and remodelling of living tissues: perspectives, challenges and opportunities. *Journal of The Royal Society Interface*, 16(157):20190233, 08 2019. ISSN 1742-5689. doi: 10.1098/rsif.2019.0233. URL https://doi.org/10.1098/rsif.2019.0233.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: two new techniques for image matching. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'77, page 659–663, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc.

Claudio Battiloro, Mauricio Tec, George Dasoulas, Michelle Audirac, Francesca Dominici, et al. E (n) equivariant topological neural networks. *arXiv preprint arXiv:2405.15429*, 2024.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Icml*, volume 2, page 4, 2021.

Blender Online Community. Blender. *Blender Foundation*, 2023. URL https://www.blender.org.

Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic faust: Registering human bodies in motion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5573–5582, 2017. doi: 10.1109/CVPR.2017.591.

Dongliang Cao, Marvin Eisenberger, Nafie El Amrani, Daniel Cremers, and Florian Bernard. Spectral meets spatial: Harmonising 3d shape matching and interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3658–3668, 2024a.

Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. Motion2vecsets: 4d latent vector set diffusion for non-rigid shape reconstruction and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20496–20506, 2024b.

Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. Grand: Graph neural diffusion. In *International conference on machine learning*, pages 1407–1418. PMLR, 2021.

Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, and Shanghang Zhang. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5291–5301, 2023.

Long Chen. Mesh smoothing schemes based on optimal delaunay triangulations. In *International Meshing Roundtable Conference*, 2004. URL https://api.semanticscholar.org/CorpusID:18536483.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Enrico Coen and Daniel J. Cosgrove. The mechanics of plant morphogenesis. *Science*, 379(6631): eade8055, 2023. doi: 10.1126/science.ade8055. URL https://www.science.org/doi/abs/10.1126/science.ade8055.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Nicolas Donati, Abhishek Sharma, and Maks Ovsjanikov. Deep geometric functional maps: Robust feature learning for shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8592–8601, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Tao Du, Kui Wu, Pingchuan Ma, Sebastien Wah, Andrew Spielberg, Daniela Rus, and Wojciech Matusik. Diffpd: Differentiable projective dynamics. *ACM Trans. Graph.*, 41(2), November 2021. ISSN 0730-0301. doi: 10.1145/3490168. URL https://doi.org/10.1145/3490168.

Efi Efrati, Eran Sharon, and Raz Kupferman. Elastic theory of unconstrained non-euclidean plates. *Journal of the Mechanics and Physics of Solids*, 57(4):762–775, 2009.

Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, and Daniel Cremers. G-msm: Unsupervised multi-shape matching with graph-based affinity priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22762–22772, 2023.

A. Sydney Gladman, Elisabetta A. Matsumoto, Ralph G. Nuzzo, L. Mahadevan, and Jennifer A. Lewis. Biomimetic 4d printing. *Nature Materials*, 2016.

Deshan Gong, Ningtao Mao, and He Wang. Bayesian differentiable physics for cloth digitalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11841–11851, 2024.

Eitan Grinspun, Anil N. Hirani, Mathieu Desbrun, and Peter Schröder. Discrete shells. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '03, page 62–67, Goslar, DEU, 2003. Eurographics Association. ISBN 1581136595.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Trans. Graph.*, 38(4), July 2019. ISSN 0730-0301. doi: 10.1145/3306346.3322959. URL https://doi.org/10.1145/3306346.3322959.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units, 2017. URL https://openreview.net/forum?id=Bk0MRI5lg.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yuanming Hu, Jiancheng Liu, Andrew Spielberg, Joshua B Tenenbaum, William T Freeman, Jiajun Wu, Daniela Rus, and Wojciech Matusik. Chainqueen: A real-time differentiable physical simulator for soft robotics. In *2019 International conference on robotics and automation (ICRA)*, pages 6265–6271. IEEE, 2019.

Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan Carr, Jonathan Ragan-Kelley, and Fredo Durand. Difftaichi: Differentiable programming for physical simulation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1eB5xSFvr.

Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W.H. Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22100–22110, 2023. doi: 10.1109/ICCV51070.2023.02025.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12627–12637, 2019.

Navami Kairanda, Marc Habermann, Christian Theobalt, and Vladislav Golyanik. Neuralclothsim: Neural deformation fields meet the thin shell theory. *Advances in Neural Information Processing Systems*, 37:110533–110572, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Jana Kierdorf, Laura Verena Junker-Frohn, Mike Delaney, Mariele Donoso Olave, Andreas Burkart, Hannah Jaenicke, Onno Muller, Uwe Rascher, and Ribana Roscher. Growliflower: An image time-series dataset for growth analysis of cauliflower. *Journal of Field Robotics*, 40(2):173–192, 2023.

Jiahui Lei and Kostas Daniilidis. Cadex: Learning canonical deformation coordinate space for dynamic surface representation via neural homeomorphism. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6624–6634, 2022.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pages 237–255. Springer, 2024a.

Yifei Li, Tao Du, Kui Wu, Jie Xu, and Wojciech Matusik. Diffcloth: Differentiable cloth simulation with dry frictional contact. *ACM Trans. Graph.*, 42(1), October 2022. ISSN 0730-0301. doi: 10.1145/3527660. URL `https://doi.org/10.1145/3527660`.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024b.

Haiyi Liang and L. Mahadevan. Growth, geometry, and mechanics of a blooming lily. *Proceedings of the National Academy of Sciences*, 108(14):5516–5521, 2011. doi: 10.1073/pnas.1007808108. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1007808108`.

Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

Simone Melzi, Jing Ren, Emanuele Rodolà, Abhishek Sharma, Peter Wonka, and Maks Ovsjanikov. Zoomout: spectral upsampling for efficient shape correspondence. *ACM Trans. Graph.*, 38(6), November 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356524. URL `https://doi.org/10.1145/3355089.3356524`.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Naoki Nishikawa, Yuichi Ike, and Kenji Yamanishi. Adaptive topological feature via persistent homology: Filtration learning for point clouds. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1TJaITmK2Q.

Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Trans. Graph.*, 31(4), July 2012. ISSN 0730-0301. doi: 10.1145/2185520.2185526. URL https://doi.org/10.1145/2185520.2185526.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International conference on learning representations*, 2020.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: revisiting pointnet++ with improved training and scaling strategies. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Gerard Salton and Michael J. McGill. Introduction to modern information retrieval. *McGraw-Hill, Inc.*, 1983.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.

David Schunck, Federico Magistri, Radu Alexandru Rosu, André Cornelißen, Nived Chebrolu, Stefan Paulus, Jens Léon, Sven Behnke, C. Stachniss, Heiner Kuhlmann, and Lasse Klingbeil. Pheno4d: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *PLoS ONE*, 16, 2021.

Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022.

Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19615–19625, 2024.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimír Vondruš, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 251–266. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/021bbc7ee20b71134d53e20206bd6feb-Paper.pdf.

Rasmus Tamstorf and Eitan Grinspun. Discrete bending forces and their jacobians. *Graph. Models*, 75(6):362–370, November 2013. ISSN 1524-0703. doi: 10.1016/j.gmod.2013.07.001. URL https://doi.org/10.1016/j.gmod.2013.07.001.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Matthew Thorpe, Tan Minh Nguyen, Hedi Xia, Thomas Strohmer, Andrea Bertozzi, Stanley Osher, and Bao Wang. GRAND++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EMxu-dzvJk.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

T. van Manen, S. Janbaz, et al. 4d printing of reconfigurable metamaterials and devices. *Communications Materials*, 2021.

Wim M. van Rees, Etienne Vouga, and L. Mahadevan. Growth patterns for shape-shifting elastic bilayers. *Proceedings of the National Academy of Sciences*, 114(44):11597–11602, 2017a. doi: 10.1073/pnas.1709025114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1709025114.

Wim M. van Rees, Etienne Vouga, and L. Mahadevan. Growth patterns for shape-shifting elastic bilayers. *Proceedings of the National Academy of Sciences*, 114(44):11597–11602, 2017b. doi: 10.1073/pnas.1709025114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1709025114.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. URL http://jmlr.org/papers/v16/vapnik15b.html.

Ingo Wald, Solomon Boulos, and Peter Shirley. Ray tracing deformable scenes using dynamic bounding volume hierarchies. *ACM Trans. Graph.*, 26(1):6–es, January 2007. ISSN 0730-0301. doi: 10.1145/1189762.1206075. URL https://doi.org/10.1145/1189762.1206075.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023.

R. Wang, C. Yuan, J. Cheng, et al. Direct 4d printing of ceramics driven by hydrogel dehydration. *Nature Communications*, 2024.

Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18816–18826, 2023.

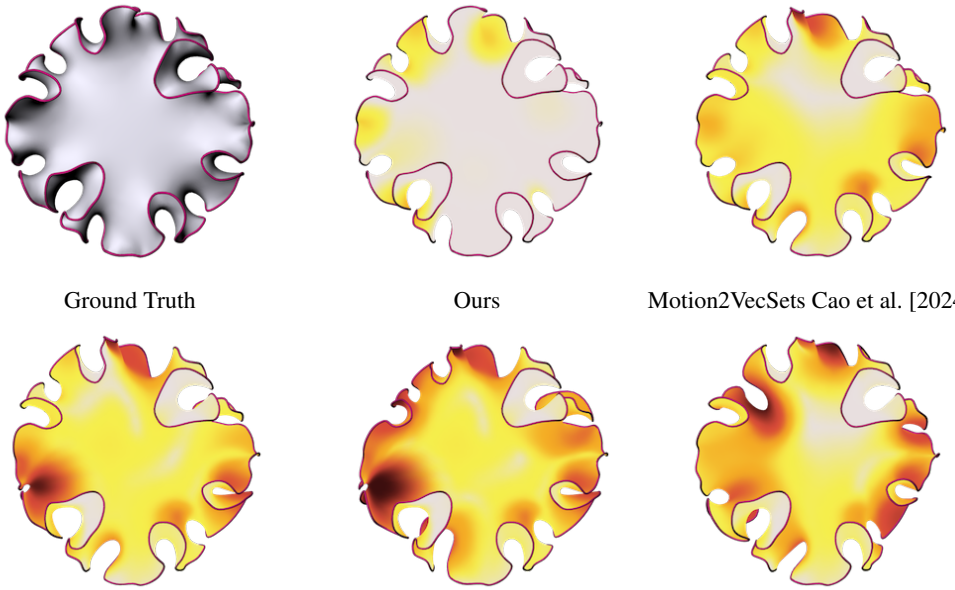
Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27091–27101, 2024.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022.

Jianzhong Zhu, Ruifang Zhai, He Ren, Kai Xie, Aobo Du, Xinwei He, Chenxi Cui, Yinghua Wang, Junli Ye, Jiashi Wang, Xue Jiang, Yulong Wang, Chenglong Huang, and Wanneng Yang. Crops3d: a diverse 3d crop dataset for realistic perception and segmentation toward agricultural applications. *Scientific Data*, 2024.

A Appendix

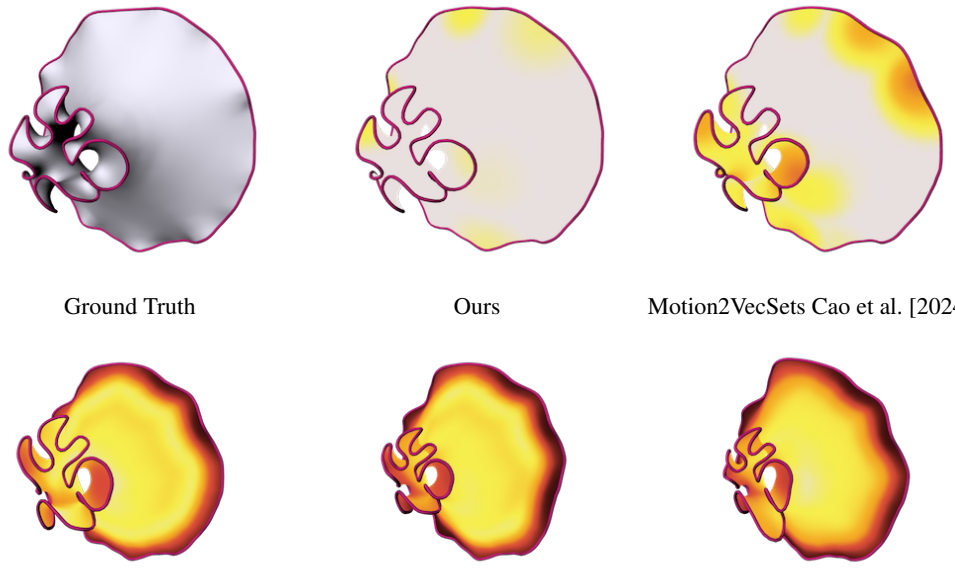
In this appendix, we provide implementation details, dataset details, and further experimental validation.



Ground Truth Ours Motion2VecSets Cao et al. [2024b]

MeshGPT-solo Siddiqui et al. [2024] CaDeX Lei and Daniilidis [2022] MeshGraphNets Pfaff et al. [2020]

Figure 4: Mesh predictions on SURF-BENCH. $\Delta t = 18$, action = $[0.1, 0.3, 0.01]$.



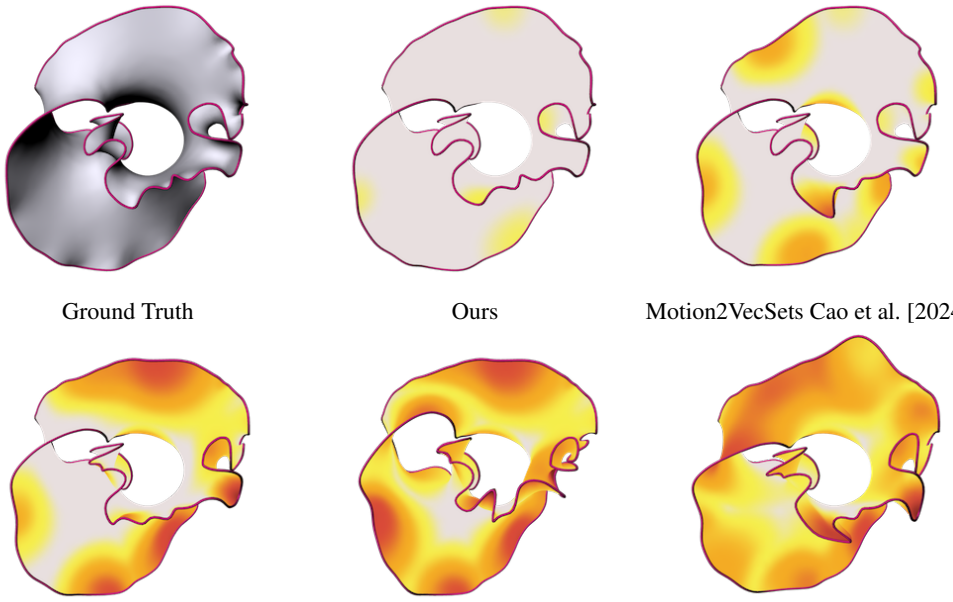
Ground Truth Ours Motion2VecSets Cao et al. [2024b]

MeshGPT-solo Siddiqui et al. [2024] CaDeX Lei and Daniilidis [2022] MeshGraphNets Pfaff et al. [2020]

Figure 5: Mesh predictions on SURF-BENCH. $\Delta t = 8$, null action.

B Sensor Tokenization and Correspondence Graph

This section provides implementation details for the geometric tokenization.



Ground Truth | Ours | Motion2VecSets Cao et al. [2024b]

MeshGPT-solo Siddiqui et al. [2024] | CaDeX Lei and Daniilidis [2022] | MeshGraphNets Pfaff et al. [2020]

Figure 6: Mesh predictions on SURF-BENCH. $\Delta t = 10$, action = $[0.05, 0.2, 0.12]$.

Shared token dimension. All modality tokens are projected to a shared dimension ($d = 768$). Unless otherwise stated, FOLIAGE uses the same token dimension for image, point, mesh, scaffold, front, boundary, action, and horizon tokens.

Image tokens. Each RGB image I_t is divided into non-overlapping 16×16 patches and encoded with a lightweight ViT Dosovitskiy et al. [2021] backbone. Patch features are linearly projected to dimension d, producing image tokens

$$T_I = \{u_k^I\}_{k=1}^{N_I}.$$

Image tokens carry a modality embedding and camera/view embedding. They do not carry intrinsic age; when image-to-mesh correspondence is available, age and growth-activity supervision are inherited from the corresponding mesh element.

Point tokens. Each point cloud P_t is downsampled to 2,048 points using farthest-point sampling. A local point transformer groups these into $N_P = 512$ point tokens,

$$T_P = \{u_k^P\}_{k=1}^{N_P}.$$

Point features include 3D position, estimated normal, local density, and relative offsets within the local point neighborhood. As with image tokens, point tokens inherit age and activity targets through nearby mesh correspondences when available. We use PointNeXt-L Qian et al. [2022] in a two-level PointNet++ Qi et al. [2017] hierarchy. Training augmentations include random point dropout, jitter, and global rotations to mimic LiDAR noise and sparsity.

Mesh tokens. The mesh $M_t = (V_t, E_t, F_t)$ can contain many more vertices than the model can process directly. We therefore form a mesh-anchor token set T_M with a budget of $N_M = 2048$. Young vertices are prioritized up to half of the budget, and the remaining anchors are selected by geodesic farthest-point sampling over the surface. Each mesh token contains position, normal, local area, curvature estimate, vertex degree, and a birth-time-derived age feature. If vertex v first appears at time $\tau(v)$, we use normalized age

$$\alpha_t(v) = \mathrm{clip}\left(\frac{t - \tau(v)}{\Delta_{\mathrm{age}}}, 0, 1\right),$$

with $\Delta_{\text{age}} = 50$, and a recency feature

$$r_t(v) = \exp\left(-\frac{\alpha_t(v)}{\tau_{\text{rec}}}\right),$$

with $\tau_{\text{rec}} = 0.2$. Age is used as an observable proxy for material maturity. We use two mesh diffusion layers (DiffusionNet Sharp et al. [2022]) to capture local geometry, followed by two learned-step graph ODEs (GRAND++ Chamberlain et al. [2021], Thorpe et al. [2022]) handles evolving connectivity.

Correspondence graph. Fused tokens interact through a sparse heterogeneous graph $G_t = (U_t, \mathcal{E}_t)$. The edge set includes:

1. **Image-to-mesh edges:** each image token connects to visible mesh elements projected into the corresponding patch. Edge features include barycentric weights, depth, visibility confidence, and view direction.
2. **Point-to-mesh edges:** each point token connects to its three nearest mesh anchors. Edge features include Euclidean distance, normal agreement, and inverse-distance confidence.
3. **Mesh-to-mesh edges:** mesh anchors connect through one-hop mesh adjacency and a small set of geodesic nearest neighbors.
4. **Within-modality local edges:** point tokens and image tokens retain local neighborhood edges within their modality.

Fusion uses type-aware sparse attention:

$$\tilde{u}_i = u_i + \sum_{j \in \mathcal{N}(i)} \text{Attn}_{\text{type}(i),\,\text{type}(j)}\big(u_i, u_j, b_{ij}\big),$$

where b_{ij} contains the edge features above. We use two sparse fusion layers in the main model.

Age propagation to non-mesh tokens. Image and point tokens inherit an age proxy from neighboring mesh anchors through correspondence edges. If multiple mesh anchors are associated with a token, we use a conservative minimum-age rule, assigning the token the youngest nearby age. Tokens without valid mesh correspondence receive a learned unknown-age embedding and are excluded from age-supervised losses.

Structured masking. During training, FOLIAGE applies three masking schemes: random token masking within each active modality, paired masking of tokens connected by correspondence edges, and modality-level dropout. We use modality dropout probability 0.3. Paired masking prevents the model from relying on trivial one-to-one correspondence copying and improves robustness when local registrations are missing or corrupted.

C Growth-Adaptive State Allocation Implementation

This section provides implementation details for GSA, the state-allocation module.

Default token counts. Unless otherwise stated, FOLIAGE uses $K_m = 128$ mature scaffold tokens, $K_f = 256$ active growth-front tokens, and $K_b = 64$ interface tokens. These values are chosen to keep the dynamics state substantially smaller than the full mesh while retaining enough active tokens for local growth.

Activity score. For each fused token u_i, the activity score is

$$g_i = \sigma\left(f_\psi\big([u_i, r_i, c_i, \rho_i, m_i]\big)\right),$$

where r_i is the recency feature, c_i contains local curvature and normal-variation features, ρ_i is average correspondence confidence, and m_i is a modality mask. f_ψ is a two-layer MLP with hidden dimension $d/2$, GELU activation, and sigmoid output. For non-mesh tokens, r_i and c_i are inherited from corresponding mesh anchors when available; otherwise learned unknown-feature embeddings are used.

Mature scaffold selection. Mature scaffold tokens are selected from low-activity or older regions. We sample K_m tokens by weighted farthest-point sampling, with sampling priority

$$w_i^{\mathrm{mat}} = (1 - g_i)(1 - r_i).$$

When mesh geodesic distances are available, farthest-point sampling uses geodesic distance; otherwise, it uses Euclidean distance in 3D. The selected scaffold tokens are refined by a graph-attention block over mesh-neighborhood and correspondence edges.

Growth-front selection. Growth-front tokens are selected as the top K_f tokens under g_i. When mesh observations are available, at least 25% of the selected front tokens are required to be mesh anchors. This prevents visually salient but geometrically unsupported image patches from dominating the active state. If fewer than K_f valid high-activity tokens are available, the front set is padded with tokens adjacent to selected front tokens in the correspondence graph.

Interface token selection. Interface tokens are selected from the transition region between active front and mature scaffold. A token is eligible if it lies within two mesh-graph hops of a selected front token and has at least one graph neighbor outside the front. Eligible tokens are ranked by

$$g_i(1 - g_i) + \eta\,\mathrm{cut}_i,$$

where cut_i counts graph edges crossing between front and non-front regions and $\eta = 0.1$. The top K_b eligible tokens form S_t^{bdry}. If the eligible set is too small, we fill the remaining slots with nearest non-front neighbors of front tokens.

Global summary. The global state vector is computed by attention pooling over all fused tokens:

$$s_t^{\mathrm{global}} = \mathrm{AttnPool}(U_t).$$

The final structured state is

$$S_t = \left\{ s_t^{\mathrm{global}}, S_t^{\mathrm{mat}}, S_t^{\mathrm{front}}, S_t^{\mathrm{bdry}} \right\}.$$

Ablation variants. We evaluate the following GSA variants:

- **Flat / homogeneous FOLIAGE:** replaces the structured state with the previous global-plus-young pooled latent.
- **Random front:** selects K_f front tokens uniformly at random while keeping the same token budget.
- **Front-only:** keeps active front tokens but removes mature scaffold and interface tokens from the dynamics state.
- **Scaffold + front, no interface:** keeps mature and front tokens but removes boundary/interface tokens.
- **Full GSA:** uses mature scaffold, active front, and interface tokens.

These variants isolate whether gains come from growth-aware selection, mature scaffold context, and front–scaffold interface modeling rather than from additional capacity alone.

Matched state-allocation controls. To separate growth-specific allocation from generic token selection, we additionally compare GSA against four matched alternatives. All variants use 448 non-global state tokens, approximately 43M parameters, the same observation encoders, geometry readout, training data, losses, and comparable transition computation. **Uniform FPS** allocates the state across the whole observed surface using farthest-point sampling, without an activity-dependent distinction between mature and future-active regions. **Learned queries** replaces geometry-selected state tokens with generic learned latent registers. **Generic pruning** uses learned adaptive token selection without the scaffold/front/interface decomposition or growth-specific activity target. **GSA + flat dynamics** retains growth-aware GSA selection but replaces the role-specific front, interface, and scaffold updates with a homogeneous transition. Full FOLIAGE combines GSA with the structured dynamics described in Sec. 3.4.

These controls distinguish three possible sources of improvement: additional state capacity, generic adaptive selection, and allocation specialized to predicted growth. Because token count and transition cost are matched, differences primarily reflect how state capacity is assigned and updated rather than how much capacity is available.

D Action-Conditioned Dynamics Details

This section provides implementation details for the dynamics module.

Action encoding. Material controls are represented as

$$a_t = [k_{\text{stretch}}, k_{\text{shear}}, k_{\text{bend}}] \in \mathbb{R}^3.$$

A two-layer MLP maps a_t to an action token

$$e_a = \text{MLP}_a(a_t) \in \mathbb{R}^d.$$

When material controls are unavailable, e_a is replaced by a learned null-action token. During training, the action token is dropped with probability 0.1 to preserve a passive prediction pathway.

Horizon encoding. The horizon Δt is encoded using Fourier features:

$$\phi(\Delta t) = \Big[\sin(2^0\pi\Delta t/T), \cos(2^0\pi\Delta t/T), \ldots, \sin(2^{L-1}\pi\Delta t/T), \cos(2^{L-1}\pi\Delta t/T)\Big].$$

We use $L = 8$ and $T = 20$. A linear projection maps $\phi(\Delta t)$ to a horizon token $e_\Delta \in \mathbb{R}^d$.

Dynamics input. The dynamics transformer receives

$$Z_t = \{s_t^{\text{global}}, e_a, e_\Delta\} \cup S_t^{\text{mat}} \cup S_t^{\text{front}} \cup S_t^{\text{bdry}}.$$

The structured update is implemented in three stages.

Front update. Growth-front tokens receive the highest-capacity update:

$$\hat{S}_{t+\Delta t}^{\text{front}} = S_t^{\text{front}} + F_\theta\left(S_t^{\text{front}}, \{s_t^{\text{global}}, e_a, e_\Delta\} \cup S_t^{\text{bdry}}\right).$$

F_θ is a two-layer Transformer block with pre-normalization, local geometric bias, and residual feed-forward updates. Front tokens attend densely to one another and to action, horizon, global, and boundary tokens.

Interface update. Interface tokens are updated after the front:

$$\hat{S}_{t+\Delta t}^{\text{bdry}} = S_t^{\text{bdry}} + B_\theta\left(S_t^{\text{bdry}}, \hat{S}_{t+\Delta t}^{\text{front}}, S_t^{\text{mat}}, s_t^{\text{global}}\right).$$

B_θ is a single cross-attention block. Interface tokens attend to predicted front tokens and nearby scaffold tokens using the graph neighborhoods defined by GSA.

Mature scaffold update. Mature scaffold tokens receive a lightweight residual update:

$$\hat{S}_{t+\Delta t}^{\text{mat}} = S_t^{\text{mat}} + \lambda_{\text{mat}}\, M_\theta\left(S_t^{\text{mat}}, \hat{S}_{t+\Delta t}^{\text{bdry}}, s_t^{\text{global}}\right),$$

with $\lambda_{\text{mat}} = 0.25$. M_θ is a one-layer graph-attention block restricted to scaffold-neighborhood edges and pooled interface context.

Global update. The global token is updated from the predicted structured state:

$$\hat{s}_{t+\Delta t}^{\text{global}} = s_t^{\text{global}} + G_\theta\left(s_t^{\text{global}}, \text{Pool}(\hat{S}_{t+\Delta t}^{\text{mat}}), \text{Pool}(\hat{S}_{t+\Delta t}^{\text{front}}), \text{Pool}(\hat{S}_{t+\Delta t}^{\text{bdry}}), e_a, e_\Delta\right).$$

Multi-horizon training and rollout. During training, Δt is sampled uniformly from $\{1, \ldots, 20\}$. For direct prediction, the model predicts $\hat{S}_{t+\Delta t}$ in one step using the horizon token. For rollout evaluation, we also recursively apply the one-step transition with $\Delta t = 1$. Direct prediction evaluates whether the state represents requested future geometry; recursive rollout evaluates stability under iteration.

Method	Retrieval ↑	Separation ↑	C40 ↓
Null action	18.6%	0.11	66.7
Uniform + action	66.8%	0.71	62.7
Action in encoder	78.2%	0.84	48.8
FOLIAGE	**90.4%**	**0.95**	**41.2**

Table 7: Counterfactual sibling-branch prediction from a shared observed prefix. Keeping the inferred current state independent of the queried action gives the strongest branch identification and separation.

Counterfactual sibling-branch evaluation. SURF-GARDEN provides sibling trajectories that share an identical observed prefix and diverge only after different material actions are applied. We use these branches to test whether the predicted state responds to the queried action rather than merely predicting a generic plausible continuation. We report the results in Tab. 7

For each branch point, **branch retrieval** asks whether the predicted future is closest to the ground-truth sibling generated by the queried action. We also report a **separation ratio**, which compares pairwise separation among predicted sibling futures with that among the corresponding ground-truth futures; a value of 1 indicates matching branch separation. C40 is Chamfer $\times 10^3$ after 40 recursive one-step transitions.

Placing the material action in the observation encoder improves over a generic action-conditioned model but remains worse than introducing the action only in the transition. This supports the intended counterfactual interface: the current state represents the observed surface, while the action selects among possible futures from that state.

E Train-Only Physical Guidance Details

This section provides implementation details for the training-only physics.

Context and target encoders. The context encoder E_{ctx} consumes only observable inputs:

$$S_t = \mathrm{GSA}(E_{\mathrm{ctx}}(x_t)).$$

The target encoder E_{tar} is used only during training and receives observable inputs plus per-vertex physical energies:

$$S^+_{t+\Delta t} = \mathrm{GSA}^+\big(E_{\mathrm{tar}}\big(x_{t+\Delta t}, e^{\mathrm{mem}}_{t+\Delta t}, e^{\mathrm{bend}}_{t+\Delta t}\big)\big)\,.$$

The target state $S^+_{t+\Delta t}$ supervises $\hat{S}_{t+\Delta t}$. The target encoder is updated as an exponential moving average of the context encoder with decay 0.998. Prediction gradients do not backpropagate through E_{tar}.

Energy-gated message passing. In the target encoder, per-vertex membrane and bending energies modulate the first mesh message-passing layer. For a message from vertex j to vertex i, we apply

$$m^+_{ij} = (1 + \gamma \tanh(W_e e_j))\, m_{ij},$$

where $e_j = [e^{\mathrm{mem}}_j, e^{\mathrm{bend}}_j]$ is the normalized physical energy vector and $\gamma = 0.5$. Energies are detached constants. The gate is bounded by the $\tanh$ nonlinearity and is applied only in the target branch.

Activity supervision. The GSA activity score is supervised using both future local change and physical energy. For token i, the target activity is

$$\tilde{g}_i = \mathrm{Norm}(0.5\,\delta_i + 0.5\,\bar{e}_i)\,,$$

where δ_i is normalized local displacement between t and $t + \Delta t$, and $\bar{e}_i$ is the normalized sum of membrane and bending energies projected to the token. $\mathrm{Norm}(\cdot)$ rescales values within each surface to $[0, 1]$. Image and point tokens inherit $\tilde{g}_i$ from neighboring mesh anchors through correspondence edges. Tokens without valid correspondence are excluded from the activity loss.

Method	Inverse-material MAE $\times 10^3 \downarrow$ No physics	Common energy loss	Matched guidance
DiffCloth	53.1	43.6	39.4
Uniform FPS-448	52.6	42.9	35.2
FOLIAGE	50.1	39.2	**32.1**
Method	Recursive C40 Chamfer $\times 10^3 \downarrow$ No physics	Common energy loss	Matched guidance
Motion2VecSets	75.4	69.0	65.8
Uniform FPS-448	72.8	67.1	62.7
FOLIAGE	54.8	47.6	**41.2**

Table 8: Equal privileged-supervision controls. Physical-energy supervision improves each architecture, especially for inverse-material inference, while FOLIAGE retains a substantial forecasting advantage when supervision is matched.

The activity loss is

$$\mathcal{L}_{\text{act}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \text{BCE}(g_i, \tilde{g}_i),$$

where Ω is the set of tokens with valid activity targets.

Energy prediction. During training, a lightweight energy head predicts normalized membrane and bending energies from context mesh tokens:

$$\hat{e}_i = h_{\text{energy}}(u_i).$$

We supervise this prediction using

$$\mathcal{L}_{\text{energy}} = \frac{1}{|\Omega_M|} \sum_{i \in \Omega_M} \text{Huber}(\hat{e}_i, e_i).$$

The energy head is discarded at deployment. Its purpose is to encourage observable features to correlate with hidden stress, not to require physical energies at inference.

Equal privileged-supervision controls. To determine whether FOLIAGE's gains arise simply because it receives privileged physical supervision during training, we provide the same physical-energy targets to comparison models under three supervision regimes in Tab. 8. **No physics** removes privileged-energy supervision. **Common energy loss** adds the same auxiliary per-vertex energy-prediction objective. **Matched guidance** additionally uses an EMA target and energy-gated target-side processing analogous to FOLIAGE. Test-time inputs are unchanged in every condition; physical energies are never provided at deployment.

Privileged physical labels are therefore useful supervision rather than an exclusive source of FOLIAGE's gains. Their effect is particularly large for hidden-material inference, as expected from the physical quantities they supervise. For recursive forecasting, architecture remains consequential even after supervision is matched: FOLIAGE reaches C40 41.2 compared with 62.7 for matched uniform allocation and 65.8 for Motion2VecSets.

Leakage controls. Privileged energies are used only in the target branch and auxiliary training losses. They are detached, bounded through gating, and never provided to the context encoder at inference. The target encoder is EMA-updated rather than directly optimized through privileged gradients from prediction. These controls preserve the intended separation between observation, action, and privileged physics.

F Training Objective, Readouts, and Complexity

This section provides details for the loss, evaluation heads, and computational cost.

Structured latent prediction loss. FOLIAGE predicts the future structured state $\hat{S}_{t+\Delta t}$ and matches it to the target state $S^{+}_{t+\Delta t}$. We use one global term and three token-set terms:

$$\mathcal{L}_{\text{pred}} = \left\| \hat{s}^{\text{global}}_{t+\Delta t} - s^{+,\text{global}}_{t+\Delta t} \right\|_2^2 + \lambda_m D\Big(\hat{S}^{\text{mat}}_{t+\Delta t}, S^{+,\text{mat}}_{t+\Delta t}\Big) + \lambda_f D\Big(\hat{S}^{\text{front}}_{t+\Delta t}, S^{+,\text{front}}_{t+\Delta t}\Big) + \lambda_b D\Big(\hat{S}^{\text{bdry}}_{t+\Delta t}, S^{+,\text{bdry}}_{t+\Delta t}\Big).$$

We use $\lambda_m = 0.5$, $\lambda_f = 1.0$, and $\lambda_b = 0.75$. The set distance is

$$D(A, B) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} \|a - b\|_2^2 + \frac{1}{|B|} \sum_{b \in B} \min_{a \in A} \|b - a\|_2^2.$$

This avoids imposing an arbitrary order on scaffold, front, or boundary tokens.

Full objective. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{act}} \mathcal{L}_{\text{act}} + \lambda_{\text{energy}} \mathcal{L}_{\text{energy}} + \lambda_{\text{vc}} \mathcal{L}_{\text{vc}}.$$

We use $\lambda_{\text{act}} = 0.05$, $\lambda_{\text{energy}} = 0.02$, and $\lambda_{\text{vc}} = 0.04$. $\mathcal{L}_{\text{vc}}$ is a variance-covariance regularizer applied to global states and pooled token states to prevent representation collapse.

Task readouts. For global tasks, we pool

$$\Big\{ \hat{s}^{\text{global}}, \text{Pool}(\hat{S}^{\text{mat}}), \text{Pool}(\hat{S}^{\text{front}}), \text{Pool}(\hat{S}^{\text{bdry}}) \Big\}$$

and apply an MLP head. This readout is used for topology classification, material regression, growth-stage recognition, and retrieval.

For dense correspondence, a token-level descriptor head predicts descriptors for mesh anchors. Descriptors are interpolated from anchors to vertices using barycentric weights, and correspondences are obtained by nearest-neighbor matching in descriptor space.

For mesh forecasting, we use a cross-attention geometry readout. Query tokens are formed from current mesh anchors plus $K_g = 256$ learned growth queries initialized near active front tokens. Each query attends to $\hat{S}_{t+\Delta t}$ and predicts a future position offset and validity score. Valid growth queries are locally triangulated with neighboring predicted anchors to reconstruct the forecast surface. This readout converts the predicted state into geometry for evaluation; the world model itself remains the state transition $S_t \mapsto \hat{S}_{t+\Delta t}$.

Common changing-support forecasting interface. Several comparison methods were originally designed for dynamics on fixed support. To compare their dynamics rather than penalize their native output parameterization, we retain each method's original dynamics backbone and attach the same variable-cardinality geometry interface used for forecasting comparisons. We report the results in Tab. 9. Queries anchored to the current support predict updated positions for existing geometry, while $K_g = 256$ learned growth queries predict candidate new vertices and validity scores. Valid candidates are reconstructed using the same local triangulation procedure.

Future vertex count, birth locations, and connectivity are withheld from every method. Each model is trained against the complete future surface and evaluated using the same Chamfer and vertex-drift protocols. Thus, the common readout standardizes how a latent prediction is converted into changing-support geometry while leaving the underlying state representation and dynamics model being compared intact.

Comparison protocol. Parent-level train/validation/test splits keep all counterfactual sibling branches from a parent trajectory in the same split. Unless an experiment explicitly varies an input or supervision source, comparison methods receive the same available observation modalities, material action and horizon inputs, training parents, augmentations, optimization epochs, and validation-trial budget. The matched allocation controls additionally equalize state-token count, parameter count, and transition computation.

These controls are cplementary to the matched state-allocation study in the main paper: the former tests whether larger task-specific baselines close the gap, while the latter isolates how a fixed FOLIAGE-sized state budget should be allocated and updated.

Method	Params	GPU-h	A↑	C↓	m↑
ETNN-L	40M	18.2	94.6	–	–
MeshGPT-L	41M	19.4	–	42.9	–
Motion2Vec	40M	18.7	–	35.6	–
ULIP-2	41M	18.9	–	–	49.1
PointCLIP	40M	18.1	–	–	50.3
FOLIAGE	43M	20.8	**98.1**	**26.4**	**61.6**

Table 9: Capacity- and training-budget-matched task baselines. A is topology accuracy, C is 5-step direct forecasting Chamfer $\times 10^3$, and m is retrieval mAP (%). Scaling the comparison models does not close the task-specific gaps.

Inference path. At deployment, FOLIAGE uses only observable inputs. Given x_t, the context encoder produces fused tokens, GSA forms S_t, and the dynamics module predicts $\hat{S}_{t+\Delta t}$ under a_t and Δt. The target encoder, physical energies, and energy heads are not used. If future observations become available, the state can be refreshed by re-encoding; otherwise, the model rolls out in latent space.

Complexity. Let $N = |U_t|$ be the number of fused input tokens and $E = |\mathcal{E}_t|$ the number of sparse correspondence edges. Token fusion costs $\mathcal{O}(Ed)$ per sparse attention layer. GSA selects K_m mature scaffold tokens, K_f active front tokens, and K_b interface tokens. Dynamics cost is dominated by front self-attention and front-interface cross-attention:

$$\mathcal{O}\big(K_f^2 d + K_b(K_f + K_m)d + K_m k d\big)\,,$$

where k is the scaffold graph neighborhood size. With default values $K_m = 128$, $K_f = 256$, and $K_b = 64$, the dynamics state is much smaller than the full mesh. GSA trades a modest amount of structured state for better allocation of modeling capacity, especially compared with full high-resolution token dynamics over all mesh anchors.

G Robustness Protocol Details

We evaluated sensor robustness under four regimes. **Rich** uses RGB, point cloud, and mesh observations. **Typical** uses RGB and point cloud observations with incomplete mesh inputs. **Sparse** uses only one geometric modality. **Noisy** adds sensor noise and random modality dropout. The same downstream probe is used across all regimes so that performance differences reflect sensing degradation rather than task-specific tuning.

For correspondence corruption, **Corr-noise** rewires 20% of pixel/point-to-mesh edges and drops 25% of remaining cross-modal edges. **Occlusion** masks 50% of RGB tokens and co-masks linked 3D tokens. **Drift** injects camera pose and intrinsics jitter before constructing cross-modal links. **Drift+SLAM** applies pose refinement before fusion. These settings test whether correspondence-aware fusion behaves as a robust grounding interface rather than an oracle dependency.

H Initialization Perturbation

We additionally stress-test cross-modal retrieval under inaccurate initial mesh alignment. We perturb the initial mesh before constructing cross-modal links by applying translation proportional to the bounding-box diagonal and a random rotation. FOLIAGE degrades monotonically as perturbation increases, but remains ahead of the correspondence-free ablation and retrieval baselines under mild and medium perturbations. Pose refinement partially recovers performance in the strong-perturbation setting. The results are report in Tab. 10..

I Long-Horizon Rollout Stability

We first evaluate recursive rollouts at horizons $(5, 10, 20, 40)$ in Tb. 11. FOLIAGE updates active growth-front tokens while preserving mature scaffold context, reducing drift accumulation over long

Init perturbation	FOLIAGE	w/o correspondence fusion	PointCLIPv2	ULIP-2
Clean	0.616	0.502	0.480	0.460
Mild: 5% bbox + 3°	0.568 ± 0.010	0.486 ± 0.009	0.456 ± 0.010	0.444 ± 0.009
Medium: 10% bbox + 6°	0.511 ± 0.012	0.463 ± 0.010	0.414 ± 0.011	0.402 ± 0.010
Strong: 20% bbox + 12°	0.445 ± 0.014	0.420 ± 0.012	0.347 ± 0.012	0.338 ± 0.011
Strong + pose refinement	0.485 ± 0.013	0.431 ± 0.011	0.361 ± 0.012	0.352 ± 0.011

Table 10: Cross-modal retrieval under initialization perturbations.

horizons. Chamfer error increases with horizon for all methods, but FOLIAGE grows more slowly than the strongest baseline.

Horizon	Best baseline Chamfer ↓	FOLIAGE Chamfer ↓
5	0.022	**0.014**
10	0.029	**0.022**
20	0.045	**0.026**
40	0.075	**0.041**

Table 11: Long-horizon rollout (recursive) stability measured by Chamfer distance.

Full post-prefix trajectories. The 40-step experiment above targets the early regime in which recursive errors begin to compound. Each SURF-GARDEN trajectory, however, contains a 50-frame common prefix followed by 350 future frames. We therefore also evaluate the complete post-prefix trajectory by recursively applying the one-step transition without re-initializing the state from ground-truth future observations. The results are reported in Tab. 12

All models accumulate substantial error over sufficiently long open-loop prediction, as expected. The ordering nevertheless remains stable throughout the trajectory: growth-aware allocation reduces error relative to homogeneous FOLIAGE, and the full scaffold/front/interface dynamics further reduce accumulation. The advantage therefore does not arise only from short-horizon prediction.

J Distributed Growth and Activity Localization

The term "growth front" denotes a functional state role rather than a geometric boundary. A future-active region may occur anywhere on the surface and may consist of multiple disconnected or interior regions. To characterize the regime in which adaptive allocation is useful, we vary the fraction of the surface undergoing substantial future change while holding topology, material distribution, sensing, and state-token budget fixed.

In Tab. 13, the absolute forecasting error rises for every model as a larger fraction of the surface changes. FOLIAGE's advantage is largest under spatially nonuniform growth, where a bounded state budget can concentrate on a comparatively small future-active region. As activity becomes global, activity recall necessarily decreases and uniform allocation becomes increasingly appropriate, causing the gap to narrow. Error analysis shows that the largest local errors occur in high-change regions omitted by the activity selector, identifying activity mislocalization as a principal failure mode of GSA.

K Transfer Protocol Details

For GrowliFlower, we evaluate growth-stage recognition using a frozen FOLIAGE encoder, a few-shot linear probe, and light fine-tuning of the last encoder block. For Pheno4D and Crops3D, correspondences are initialized using anchored ICP with pruning and temporal pose smoothing before applying learned descriptors. For D-FAUST and CAPE, we use the standard correspondence structure provided by the datasets and evaluate whether the learned descriptors transfer to dynamic 3D surfaces beyond the growing-surface simulator.

These protocols test external validity without relying on simulator oracle correspondences. The controlled simulator provides counterfactual supervision and hidden physical quantities; transfer

Method	C40↓	C100↓	C200↓	C350↓
Motion2VecSets	75.4	158.2	293.6	489.1
Uniform FPS-448 AR	62.7	133.2	250.9	415.1
FOLIAGE w/o GSA	52.7	103.1	190.2	319.6
FOLIAGE	**41.2**	**82.4**	**153.7**	**261.8**

Table 12: Full post-prefix recursive rollout. Values are Chamfer $\times 10^3$. Absolute error grows for every model, but FOLIAGE accumulates error more slowly over the complete 350-frame future.

Method	Localized 10% C40↓	Diffuse 50% C40↓	Global 100% C40↓
Uniform FPS	62.7 ± 1.8	76.9 ± 2.1	107.8 ± 3.1
FOLIAGE w/o GSA	52.7 ± 1.4	72.1 ± 2.0	106.3 ± 3.0
FOLIAGE	$\mathbf{41.2 \pm 1.1}$	$\mathbf{61.9 \pm 1.7}$	$\mathbf{103.6 \pm 2.9}$
Activity Recall@256	0.924	0.712	0.512

Table 13: Effect of increasingly distributed growth. C40 is recursive Chamfer $\times 10^3$; mean $\pm$ standard deviation is over five independently trained seeds. Activity Recall@256 measures coverage of future-active surface by the 256 selected front tokens.

experiments test whether the learned representation remains useful when those simulator-specific signals are absent.

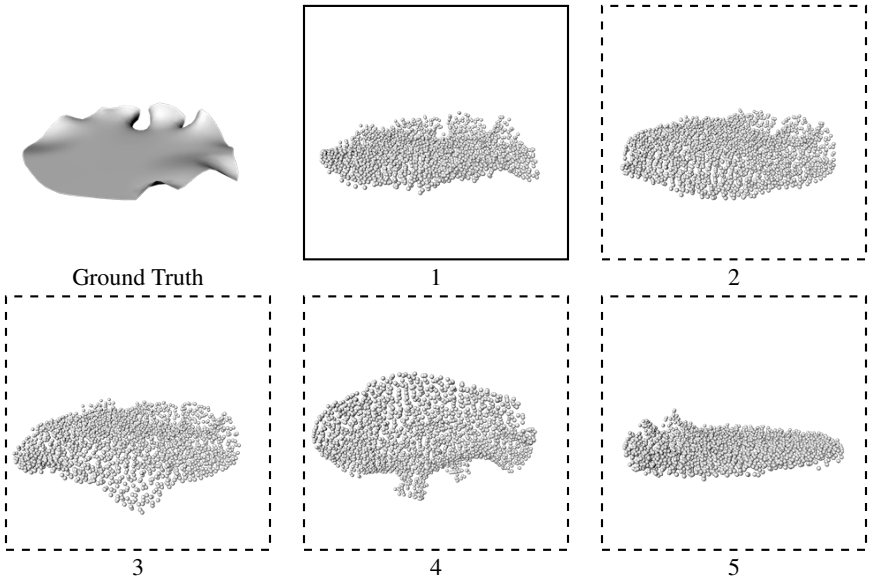


Figure 7: Top-5 retrievals on SURF-BENCH (Image → Point Cloud)

Passive real-world forecasting protocol. We use the same baselines and report the resulting forecast errors in Tab. 6. Chamfer distance is normalized by the plant bounding box and reported as $\times 10^3$. Real forecasting results average three fine-tuning seeds; confidence intervals are 95% bootstrap intervals over held-out plant trajectories.

The two-week cabbage setting contains large changes in plant size, leaf pose, and newly observed surface. FOLIAGE obtains newly observed-surface F1 of 55.3% [53.1,57.5], compared with 28.7% [26.6,30.8] for persistence. This test therefore requires more than retaining the geometry already visible at the input time. At the same time, these datasets provide only naturally occurring passive growth: SURF-GARDEN remains complementary because it provides matched initial states under alternative material interventions.

Latent Dim. d	**Score**	**EMA Rate**	**Score**	**Sampling Range** (Δt)	**Score**
512	0.74	0.995	0.80	Uniform 1–4	0.78
640	0.79	0.997	0.81	Uniform 1–12	0.80
768 (Ours)	**0.82**	0.998 (Ours)	**0.82**	Uniform 1–20 (Ours)	**0.82**
896	0.81	0.999	0.81	Uniform 1–28	0.81
1024	0.78	0.9995	0.79	Uniform 1–36	0.79

Table 14: Ablation results for model capacity and temporal encoding. Each block shows the effect of sweeping a single hyperparameter on the composite validation score.

Method	Material MAE $\times 10^3 \downarrow$	Direct C5 $\times 10^3 \downarrow$	Recursive C40 $\times 10^3 \downarrow$
DiffCloth	53.1 ± 1.1 [52.2,54.0]	–	–
Motion2VecSets	–	38.1 ± 0.7 [37.5,38.7]	75.4 ± 2.2 [73.5,77.3]
Uniform FPS-448	35.2 ± 0.8 [34.6,35.9]	28.2 ± 0.5 [27.8,28.7]	62.7 ± 1.8 [61.2,64.3]
FOLIAGE w/o GSA	35.0 ± 0.7 [34.4,35.6]	28.0 ± 0.5 [27.6,28.4]	52.7 ± 1.4 [51.5,54.0]
FOLIAGE	$\mathbf{32.1 \pm 0.6}$ [31.6,32.7]	$\mathbf{25.7 \pm 0.4}$ [25.3,26.1]	$\mathbf{41.2 \pm 1.1}$ [40.3,42.2]

Table 15: Statistical reliability of the main hidden-material and forecasting results. C5 is direct prediction at five steps; C40 is a recursive 40-step rollout.

L SURF-GARDEN Details

In SURF-GARDEN, surfaces undergo *metric accretion*, where new material is added, and deform like thin elastic shells. Energy-based simulations produce FEM-quality meshes with two sensor projections per frame. With rich physics, precise cross-modal alignment, and topological diversity, SURF-GARDEN offers a well-rounded addition to accretive surface growth research.

L.1 SURF-GARDEN

Energy Model. Formally, SURF-GARDEN embeds a non-Euclidean metric into $\mathbb{R}^3$, inducing negative curvature and curling, an approach grounded in differential geometry and plant morphogenesis research. For a mesh $\mathcal{M}_t = (V_t, E_t, F_t)$, the simulator minimizes a smooth energy Grinspun et al. [2003]:

$$\mathcal{E}(V_t) = \underbrace{k_{\text{stretch}} \sum_e (\|e\| - \ell_e^\star)^2 + k_{\text{shear}} \sum_f \|S_f\|_F^2}_{\mathcal{E}_{\text{membrane}}} + \underbrace{k_{\text{bend}} \sum_{(f_i,f_j)} (\theta_{ij} - \theta_{ij}^\star)^2}_{\mathcal{E}_{\text{flexural}}},$$

where $\ell_e^\star$ is the rest length, S_f the shear tensor, and $\theta_{ij}^\star = 0$ the preferred dihedral. The closed-form gradients Tamstorf and Grinspun [2013] are efficient to evaluate and support stable forward Euler steps at $\Delta t = 10^{-2}$.

Metric Accretion and Non-Euclidean Growth. Growth is modulated by $g(v) \in [0, 1]$, the normalized geodesic distance to a source set, simulating hormone diffusion (e.g., auxin). Rest lengths update as $\ell_e^\star(t) = ((g(v_i) + g(v_j))/2 + 1)\|e\|$. Edges that exceed $1.5\times$ their original rest length are split. This guided adaptive refinement varies growth spatially in a differential manner, with some parts growing faster than others. Coefficients $k_{\text{stretch}}, k_{\text{shear}}, k_{\text{bend}}$ introduce temporal variation.

Mesh Quality. Meshes are optimized with ODT smoothing Chen [2004] and Delaunay edge flips, producing near-isotropic triangles (0.88 ± 0.04 radius-edge ratio) and valence $5{-}7$. Self-intersections are avoided using ellipsoidal vertex colliders in a bounding-volume hierarchy (BVH) Wald et al. [2007].

Topological Variety. We include six classes of 2-manifolds with boundary: disc, annulus, punctured torus, Möbius strip, pair-of-pants, and thrice-punctured disc. These cover lobed, twisted, and compound forms common in botany and geometry alike, reducing overfitting to select shapes.

Counterfactual Branching. Each sequence begins with a 50-frame prefix under baseline parameters θ^A, after which it branches into two or more trajectories with modified controls $\theta^B, \theta^C, \ldots$ (e.g., halved k_{bend}). Identifiers in the half-edge structure ensure that vertex indices and geometry remain aligned post-branch. This enables supervised counterfactual reasoning: identical pasts yield distinct futures, and the model must predict each outcome conditioned on an action token.

Such branch-point supervision equips physically intelligent agents with the ability to forecast the consequences of their own interventions. During training, FOLIAGE observes the prefix through the context encoder and rolls out to each branch using its corresponding action tokens. At inference, alternate futures can be queried by swapping tokens—no simulator calls are required.

SURF-GARDEN provides $7,200$ branched growth sequences, each defined by $k_{\text{stretch}}, k_{\text{shear}}, k_{\text{bend}}$, a topology class, and a random seed. Every sequence spans 400 frames, evolving from rest to maturity with vertex counts that increase from 20 to 10^5. Variation is further introduced through quaternion perturbations and vector fields. We split the dataset $8:1:1$ into train/val/test.

L.2 Dataset correspondence and annotation

Each frame yields two mesh-tied modalities. *(1) Multi-view RGB*: eight cameras on a Fibonacci sphere render photorealistic frames with Cycles shading Blender Online Community [2023], HDR lighting, and 50mm lenses. Exposure jitter, defocus, and 20% CutOut masking improve robustness DeVries and Taylor [2017]. Each pixel carries its emitting triangle index and barycentric coordinates. Cameras are fixed per trajectory; masks persist over time. *(2) LiDAR-style point cloud*: a 64×2048 raycast with $\sigma{=}5$ mm Gaussian noise and 5% dropout mimics real-world sparsity. Each point stores its nearest mesh vertex. Both views preserve consistent token indices, enabling direct cross-modal supervision.

Across frames, a half-edge data structure is maintained with a unique identifier for the vertices, edges, and faces. Even as vertices and edges are added (new vertices, edge flips etc.) and their indices are updated, these identifiers remain the same, allowing us to exactly identify the same mesh elements over time alongside their quantities of interest (energy, age, etc.).

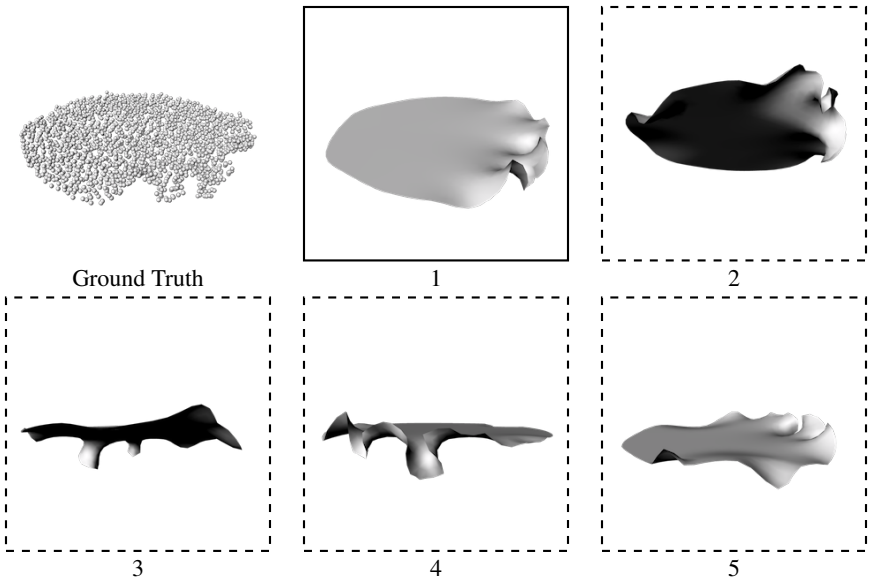


Figure 8: Top-5 retrievals on SURF-BENCH (Point Cloud → Image)

M Critic Heads

For each SURF-BENCH task, we attach a specialized critic to evaluate the core objective under realistic sensor constraints. The topology critic ingests mesh-only latent state into a frozen backbone plus a 768→256→6 MLP for genus classification; the material critic uses a single RGB-based latent with a one-hidden-layer regressor to predict bending modulus; the stage critic processes four latent state embeddings through a 128-unit Bi-GRU Cho et al. [2014] for balanced growth-stage accuracy; the growth critic conditions MeshGPT's Siddiqui et al. [2024] autoregressive split/offset tokens on M_t and $s_{t+\Delta t}$ to measure Chamfer Barrow et al. [1977] and vertex-count errors; the retrieval critic ranks image-to-mesh latent states by cosine similarity Salton and McGill [1983]; and the correspondence critic refines per-vertex features with global and young-region tokens via a 2-layer residual MLP,

projects onto 128 spectral components, solves a functional map Ovsjanikov et al. [2012], and applies five ZoomOut refinements Melzi et al. [2019].

N Action Encoder

The action space in our unbounded surface evolution settings consists of the three scalar elastic coefficients that parameterise the shell mechanics, $\mathbf{a}_t^{\text{raw}} = [k_{\text{stretch}},\, k_{\text{shear}},\, k_{\text{bend}}]$. Because these values span several orders of magnitude, we first apply a logarithmic re–scaling $\tilde{k} = \log_{10}(k),\ k \in \{k_{\text{stretch}}, k_{\text{shear}}, k_{\text{bend}}\}$ followed by z–score normalization using the mean and variance estimated throughout the training set. The normalized vector $\tilde{\mathbf{a}}_t \in \mathbb{R}^3$ is then embedded into the model's token space through a two–layer perceptron $\mathbf{a}_t = \text{MLP}_{\text{act}}(\tilde{\mathbf{a}}_t), \text{MLP}_{\text{act}} :\ 3 \to 128 \to d,\ d = 768$ with GELU activations Hendrycks and Gimpel [2017] and layer normalization Ba et al. [2016]. This produces the action token $\mathbf{a}_t \in \mathbb{R}^{768}$.

At training time, the encoder also encounters a learned $\mathbf{a}_{\text{null}}$ embedding that substitutes for $\mathbf{a}_t$ whenever material coefficients are withheld. During training we drop the entire action token with probability 0.1 for this scenario. During inference, the user may supply a physical–coefficient vector to perform counterfactual roll–outs; if omitted, the encoder inserts the null token, reverting the model to passive prediction behavior.

O Mesh Forecasting

Fig. 4, 5, and 6 shows examples of mesh prediction by FOLIAGE and baselines across different look-ahead and action conditioning. For simple topology and limited growth, the general morphology of the surface is preserved. But near the boundaries and in areas of high feature activity (emergence or disappearing of buckling), prediction error increases, especially for baseline models. In Fig. 5, baselines such as MeshGraphNets Pfaff et al. [2020] and CaDeX Lei and Daniilidis [2022] struggle to model surfaces that had enlarged substantially through accretive growth, resulting in visibly 'shrunk-down' predictions. With more complex topology such as the Möbius strip (Fig. 6), these errors propagate globally. These disparities highlight the effectiveness of FOLIAGE's perception-action setup and physics-guided learning to model the complex deformations and growth of the surfaces.

P Cross-Model Retrieval

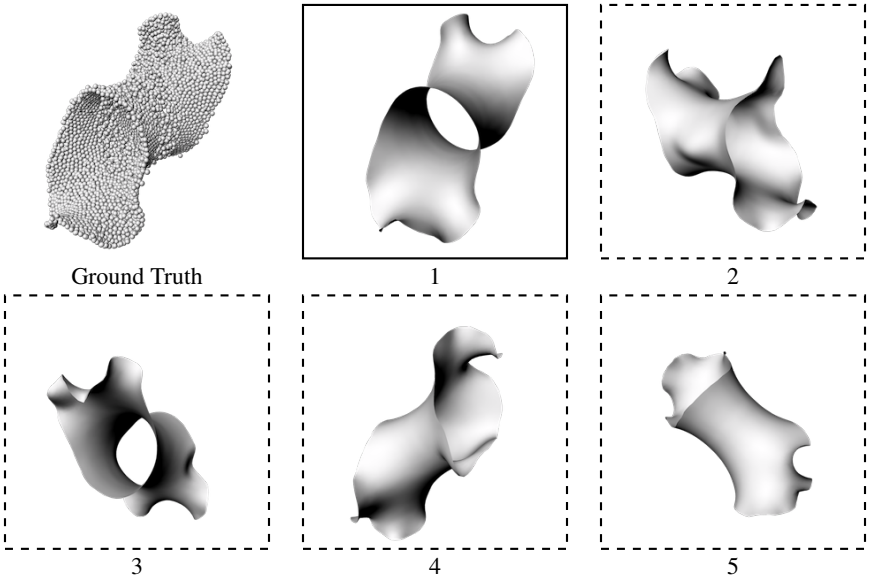


Figure 9: Top-5 retrievals on SURF-BENCH (Point Cloud → Image, Zero-shot)

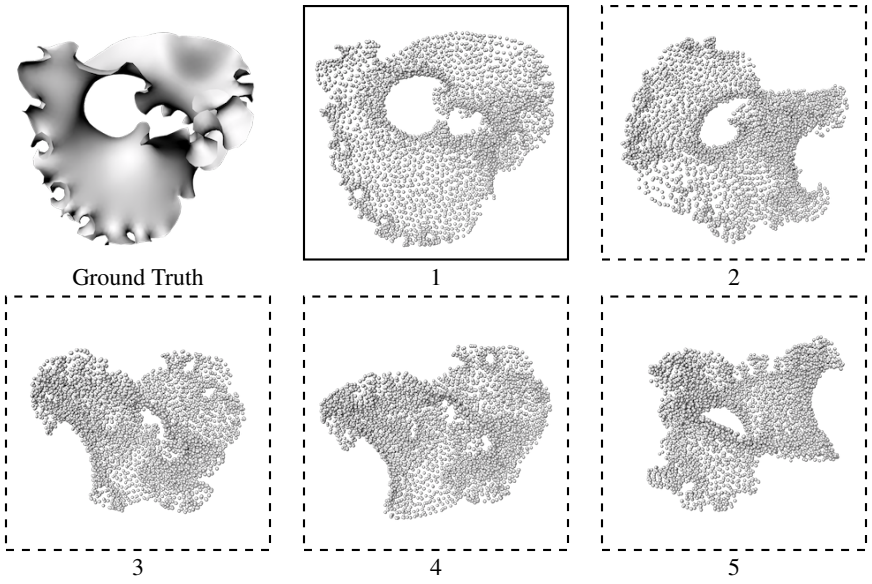


Figure 10: Top-5 retrievals on SURF-BENCH (Image → Point Cloud, Zero-shot)

In Fig. 7, 8, 9, and 10 we show examples of cross-model retrieval in normal and zero-shot settings for point clouds and images. The correct option (solid line border) is differentiated from the incorrect ones (dashed line border). We note that the purple lines highlighting the boundary (e.g. in Fig. 4) are a visual aid; they are not present in the rendered images of the mesh surface. FOLIAGE's first choice is predominantly the correct one followed by visually similar surfaces (image rendering or point cloud representations) with the same topology categorization. This is observed for unseen examples with complex morphology and challenging viewing angles, indicating the strong semantic awareness and consistency of FOLIAGE's latent state across different modalities.

Q Dense Correspondence

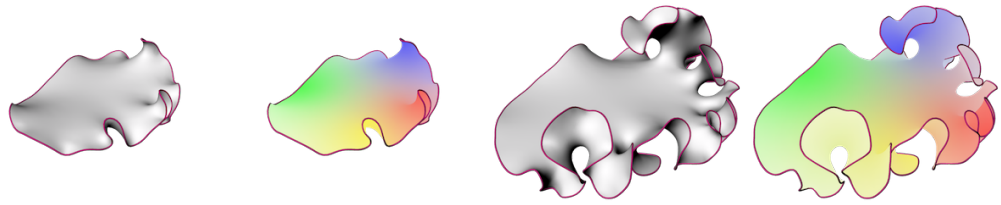


mesh at $t = 100$ (b) mesh at $t = 108$ (FOLIAGE)

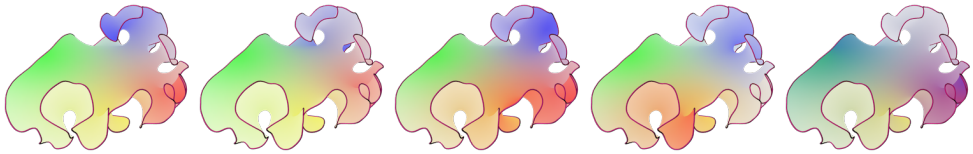
Figure 11: A vertex feature (turquoise sphere) that began at the bottom of a valley (left) quickly develops into the peak of a mountain (right).

In contrast to well-studied domains like human or animal bodies, which follow a set template (the skeleton) movements constrained to specific parts of the geometry (e.g. arm movements has very limited impact on the full body), SURF-GARDEN's open surfaces represent a continuum in which features smoothly emerge and dissipate. As Fig. 11 suggests, identifying one of a fixed number of extrusions (e.g. fingers on a hand) is insufficient in the accretive growth regime.

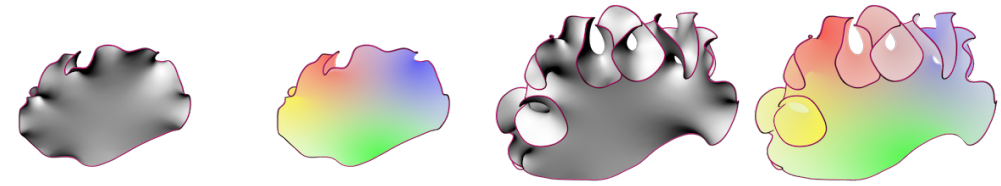
Fig. 12, 13, and 14 show the correspondence maps from a source mesh to a target mesh generated by FOLIAGE and baselines from two distinct viewing angles each. As the surface expands and buckles, a small bulge quickly develops into multiple twists and turns which FOLIAGE reliably tracks. Meanwhile, baseline models increasingly lose track of or mismatches features as the morphological complexity of the surface grows under shell physics and material accretion.



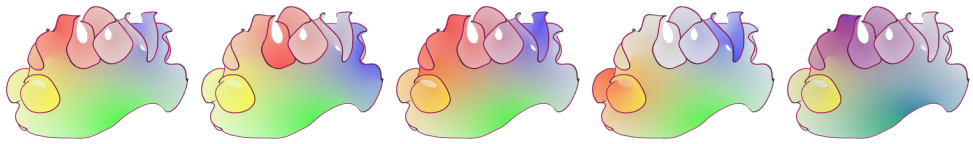
Source view A Target (GT) view A



Ours Cao et al. [2024a] DiffusionNet G-MSM ZoomOut



Source view B Target (GT) view B



Ours Cao et al. [2024a] DiffusionNet G-MSM ZoomOut

Figure 12: Correspondences on SURF-BENCH.

R SURF-GARDEN Parameters

SURF-GARDEN supports the exploration of a large morphology space guided by physical control parameters $k_{\text{stretch}}, k_{\text{shear}}, k_{\text{bend}}$. Fig. 15 illustrates the effect of their different combinations. A low bending coefficient models a thinner, more flexible surface that is prone to more complex deformations; a higher value models a thicker surface that only permits large-scale deformations to form slowly. Stretching and shearing coefficients further regulate the local behavior of the surface, leading to varied morphology.

S Latent Space

Topology. FOLIAGE's latent space naturally arranges shapes according to their global invariants—genus and boundary count—while not explicitly trained on topology classification tasks. In Fig. 16, all genus-0 surfaces (disc, annulus, pair-of-pants, thrice-punctured disc) form one region,

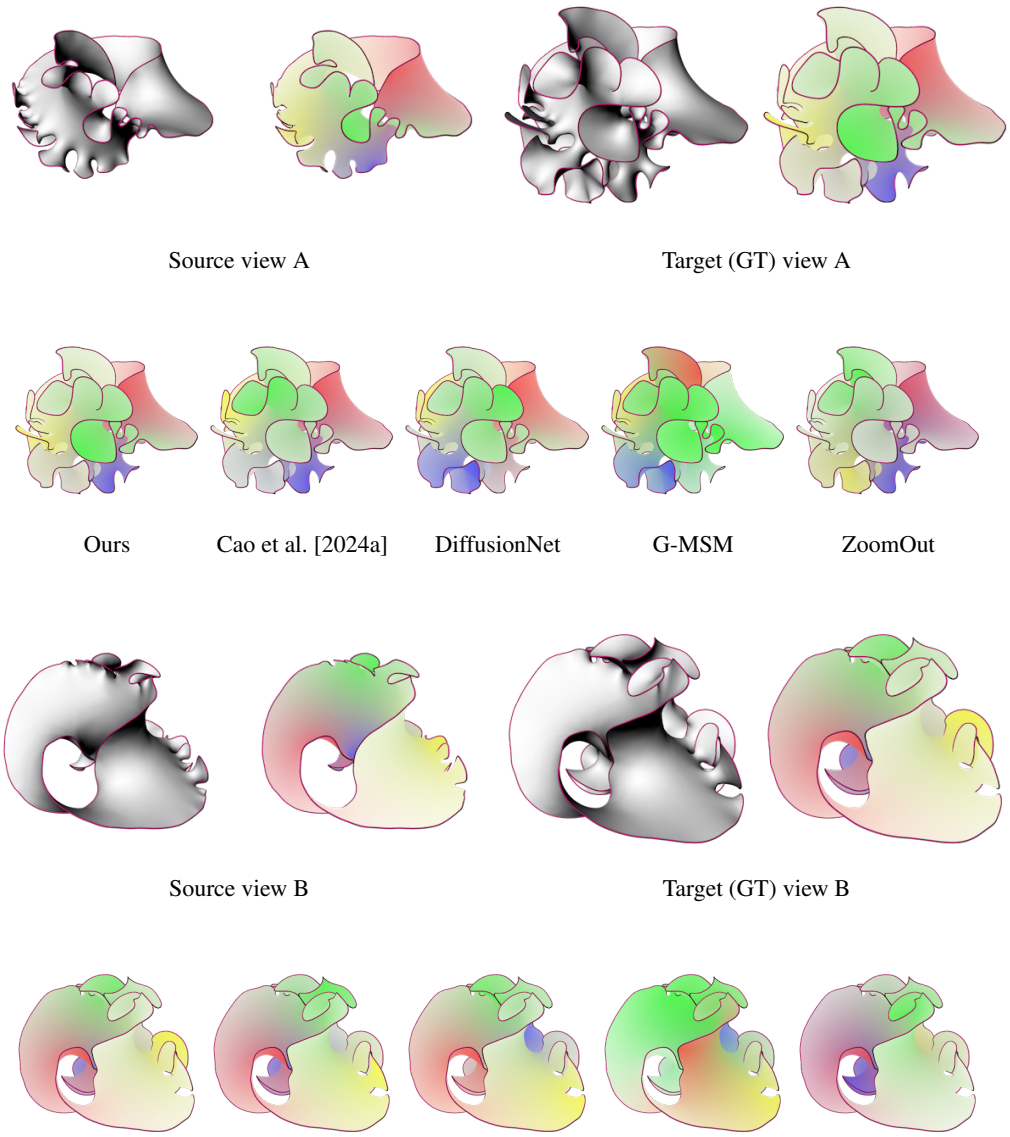


Figure 13: Correspondences on SURF-BENCH.

with boundary-count differences causing small shifts along a shared axis: for example, the disc (one boundary) sits between the annulus (two holes) and the pair-of-pants (three holes). By contrast, the genus-1 punctured torus and the non-orientable Möbius strip form a distinct cluster, reflecting their additional "handle" or "twist."

T Extended Ablation Studies

Before delving into detailed ablations, we clarify our composite-metric proxy. Rather than tuning four hyperparameters across six individual tasks (and four stress tests), we normalize each task's evaluation metric into a $[0, 1]$ range (inverting distances so that higher score indicates better performance), weight all tasks equally, and sum them into a single scalar. This composite score strongly correlates with the full multi-task performance of interest ($\rho \approx 0.92$), enabling broad five-point sweeps to be run efficiently. Once top-performing settings emerge, we re-evaluate the individual metrics for each task and report them in Tab. 14, 16, and 17.

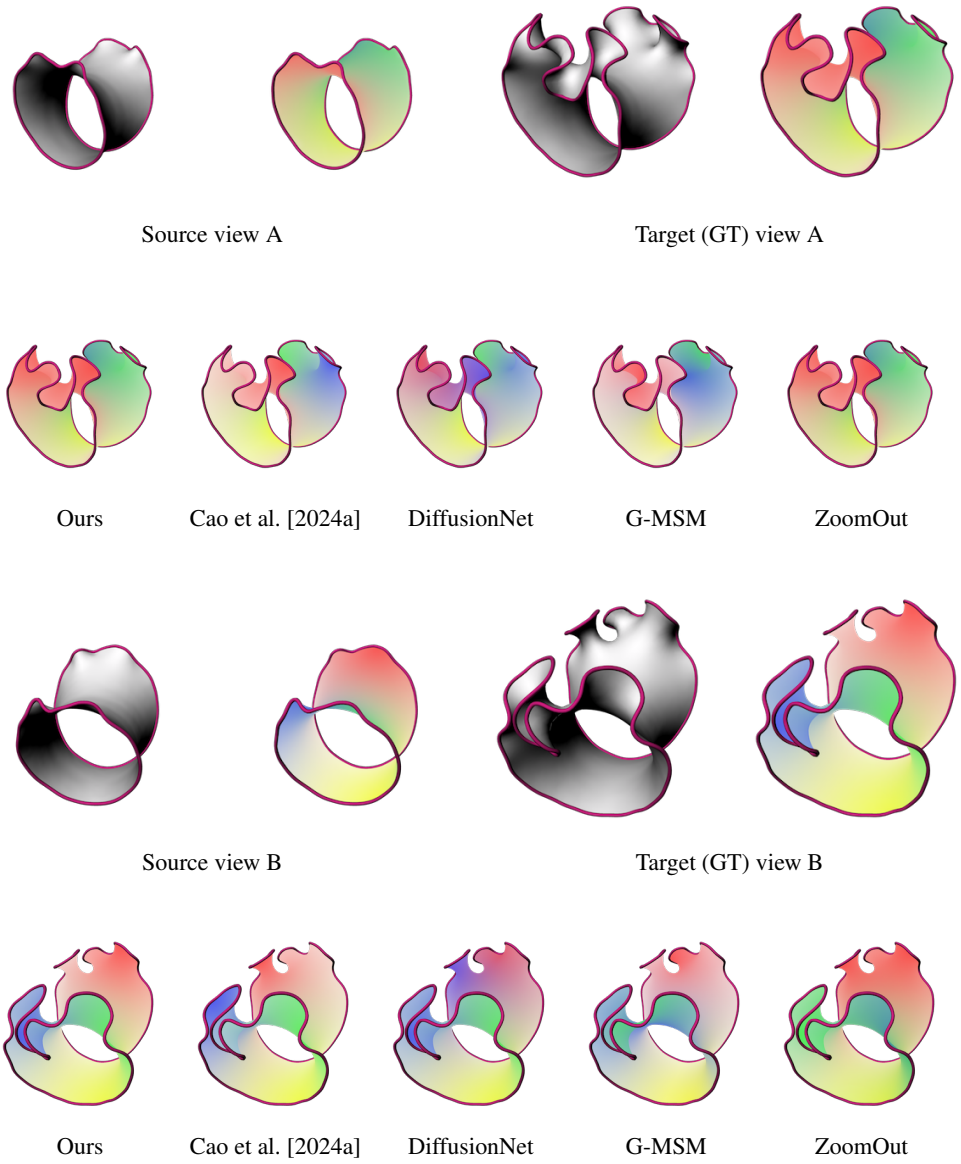


Figure 14: Correspondences on SURF-BENCH.

T.1 Statistical Reliability Across Seeds and Topology

For the experiments most directly supporting hidden-material inference and forecasting, we additionally train five models from independent random initializations. In Tab. 15, we report mean $\pm$ standard deviation across trained models; bracketed values are 95% bootstrap confidence intervals over held-out parent trajectories after averaging predictions across seeds.

The ranking is unchanged across all six SURF-GARDEN topology families. FOLIAGE's C40 ranges from 34.3 ± 0.9 on discs to 52.0 ± 1.5 on Möbius surfaces, whereas uniform allocation ranges from 48.0 ± 1.3 to 82.4 ± 2.3. On the highest-error Möbius family, the respective 95% intervals are [50.7,53.3] and [80.5,84.4]. Thus, the aggregate forecasting gain is not driven by a single topology family or training seed.

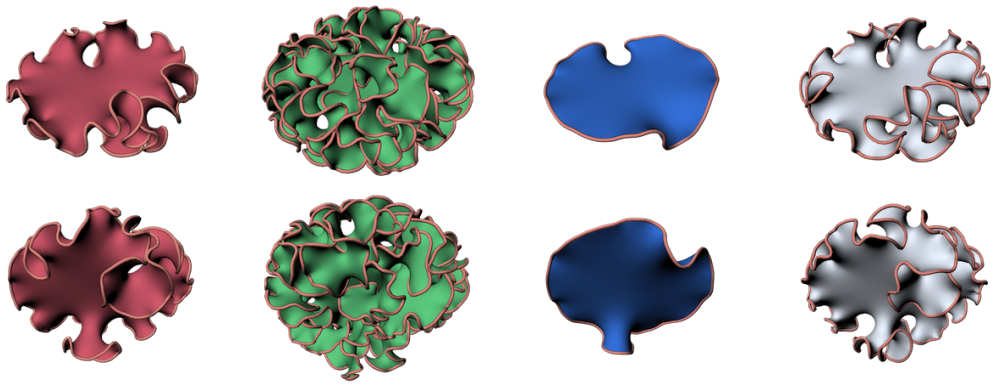


Figure 15: Effect of SURF-GARDEN physical control parameters $[k_{\text{stretch}}, k_{\text{shear}}, k_{\text{bend}}]$ (left to right columns): $[0.15, 0.15, 0.25]$, $[0.15, 0.15, 0.2]$, $[0.1, 0.15, 0.2]$, and $[0.15, 0.1, 0.2]$, respectively.

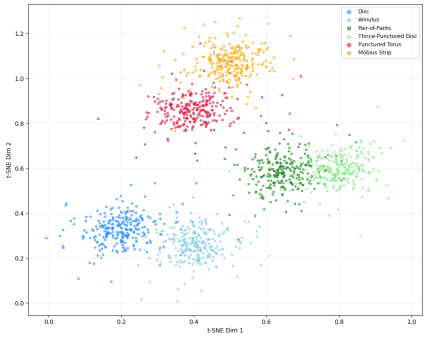


Figure 16: TSNE on topology classes.

T.2 Model Capacity and Temporal Encoding

In Tab. 14, we swept the latent dimensionality d from 512 to 1024. Smaller dimensions (512–640) consistently underperform: the model lacks sufficient capacity to encode fine-grained geometric and energetic signals, impairing tasks such as dense correspondence and material regression. Larger dimensions (896–1024) offer diminishing returns—more parameters than data—and exhibit slightly reduced stability during long-horizon rollouts, as the predictor transformer struggles to regularize across a wider channel space. We find $d = 768$ to be the optimal trade-off, balancing expressivity for physics-informed features (e.g., membrane and flexural energies) with trainability.

We also tuned the EMA (exponential moving average) update rate for the target (privileged-signal) encoder. Slower rates (0.995–0.997) update too sluggishly, causing the context and target embeddings to drift apart, which diminishes the effectiveness of energy-gated message passing. Faster rates (0.999–0.9995) over-smooth the target, preventing it from reflecting the latest context weights, and thereby hamper auxiliary energy regression. We observed that a rate of 0.998 best balances stability and responsiveness.

T.3 Temporal Horizon Sampling

Choosing the rollout-horizon distribution is critical in a physics-aware world model. As Tab. 14 shows, when we bias sampling toward short horizons (e.g., Uniform(1, 4)), the model learns only incremental dynamics and performs poorly in mid-term predictions; Chamfer and vertex-drift errors spike after 10 steps. Conversely, sampling very long horizons (e.g., Uniform(1, 36)) spreads the model's capacity across a wide temporal range, weakening both short-term fidelity and long-term coherence. Our Uniform(1, 20) policy emphasizes the early and mid-growth phases—where prediction is most critical—while still exposing the model to challenging, longer-range rollouts. Empirically, we find that this sampling regime consistently maximizes the composite score without overfitting to either extreme.

Token Drop Ratio	Score	Modality Drop Ratio	Score	Action Drop Prob.	Score
15%	0.78	20%	0.79	0%	0.80
20%	0.80	25%	0.81	5%	0.81
25% (Ours)	**0.82**	30% (Ours)	**0.82**	10% (Ours)	**0.82**
30%	0.80	35%	0.79	15%	0.81
35%	0.77	40%	0.75	20%	0.78

Table 16: Ablation results for regularization strategies. Each column group shows the effect of sweeping one dropout-related hyperparameter on the composite validation score. The selected configuration for each is highlighted in bold.

Learning Rate	Score	Weight Decay	Score	λ_{energy}	Score	λ_{vc}	Score
5.0×10^{-4}	0.78	5.0×10^{-3}	0.79	0.00	0.75	0.00	0.76
7.5×10^{-4}	0.80	7.5×10^{-3}	0.81	0.01	0.79	0.02	0.80
1.0×10^{-3} (Ours)	**0.82**	1.0×10^{-2} (Ours)	**0.82**	0.02 (Ours)	**0.82**	0.04 (Ours)	**0.82**
1.5×10^{-3}	0.79	1.5×10^{-2}	0.80	0.04	0.80	0.06	0.81
2.0×10^{-3}	0.75	2.0×10^{-2}	0.76	0.08	0.78	0.08	0.77

Table 17: Ablation results for optimization and loss weighting. Each group shows a sweep over one hyperparameter and its effect on the composite validation score.

T.4 Regularization and Robustness

Token and Modality Dropout. In our cross-modal fusion setup, we independently drop 25% of tokens per modality and 30% of entire modalities. As Tb. 16 shows, lower dropout rates (15–20% token, 20–25% modality) fail to regularize adequately: the model overfits to specific sensor patterns and degrades under simulated sensor dropout (Stress S1). Higher rates (30–35% token, 35–40% modality) deprive the fusion transformer of coherent correspondence signals, weakening geometry–correspondence alignment and degrading performance on tasks such as topology classification and retrieval. The selected dropout rates strike a balance, simulating realistic sensor failures without removing so much information that cross-modal attention cannot reconstruct object structure.

Action Dropout. We further experimented with dropping the action token during training (0%–20%) in Tab. 16. Omitting action dropout leads to a model that is overly dependent on control inputs and generalizes poorly when such inputs are noisy or absent. Conversely, excessive dropout (15–20%) enforces robustness at the cost of physical consistency, as the model may ignore legitimate control signals. A moderate 10% dropout encourages the model to infer actions from observed state transitions, while still forming tight action–perception loops when control signals are present.

U Optimization

Learning Rate and Weight Decay. As Tab. 17 shows, a low AdamW learning rate (e.g., 5×10^{-4}) leads to slow convergence and under-optimized parameters, while a high learning rate (e.g., 2×10^{-3}) causes unstable gradients, especially in the multi-head self-attention layers of the predictor. Similarly, a weak weight decay (e.g., 5×10^{-3}) under-regularizes the high-dimensional latent space, whereas overly strong decay (e.g., 2×10^{-2}) suppresses meaningful emergent physics representations. Our chosen configuration—learning rate of 1×10^{-3} and weight decay of 1×10^{-2}—yields smooth optimization and robust generalization.

Energy and Variance–Covariance Loss Weights. The auxiliary energy regression weight λ_{energy} and variance–covariance regularizer λ_{vc} govern how much the model prioritizes privileged physical signals over raw rollout accuracy. Setting the energy weight to zero $\lambda_{\mathcal{E}} = 0$ causes material inference to degrade, while excessive weight (e.g., $\lambda_{\text{energy}} = 0.08$) pulls the latent space toward physics features at the cost of open-loop prediction accuracy, worsening Chamfer and drift metrics. We select $\lambda_{\text{energy}} = 0.02$ and $\lambda_{\text{vc}} = 0.04$ to ensure that physics cues meaningfully inform the representation without overwhelming the primary learning signal, striking a balance between interpretability and predictive performance.