4DGS-JEPA: TEMPORALLY COMPOSITIONAL JOINT-EMBEDDING PREDICTION FOR DYNAMIC GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic Gaussian Splatting explicitly represents evolving 3D scenes, but is typically optimized for reconstruction, rendering, or future-state generation rather than reusable predictive dynamics. We propose 4DGS-JEPA, a Gaussian-native joint-embedding predictive architecture for causal multi-horizon prediction. It combines hierarchical scene, motion-group, and Gaussian-level representations with a horizon-conditioned transition supporting both direct prediction and recursive rollout. Its central principle is temporal composition: different chronological transition paths reaching the same future endpoint should produce compatible predictive states. Endpoint and multi-horizon path supervision anchor predictions to future target embeddings, while a selective geometry decoder and geometry-level composition ground the learned dynamics in coherent group motion and Gaussian geometry without reconstructing complete future appearance. We further introduce hybrid correspondence that preserves reliable persistent canonical identity and applies residual optimal transport when correspondence becomes ambiguous under reordering or topology change. We theoretically characterize zero-loss path agreement and finite-error rollout accumulation. In controlled dynamic Gaussian worlds, temporal composition reduces direct–composed latent discrepancy by about 61% while slightly improving predictive accuracy, geometry-level composition reduces group-motion path discrepancy by about 72%, hybrid correspondence remains robust under reordering and topology change, and predicted Gaussian geometry can be rendered into held-out future frames without pixel-level training. These results establish 4DGS-JEPA as a predictive, temporally compositional formulation of dynamic Gaussian worlds.

1 INTRODUCTION

3D Gaussian Splatting (3DGS) is a powerful explicit scene representation that combines differentiable optimization, high-quality novel-view synthesis, and real-time rendering (Kerbl et al., 2023), and dynamic extensions such as 4D-GS model time-dependent deformation of evolving scenes (Wu et al., 2024). These methods are optimized to reconstruct or interpolate an observed sequence, so accurate reconstruction at arbitrary timestamps does not by itself provide a *reusable predictive state* that can be rolled forward beyond the observations. Joint-Embedding Predictive Architectures (JEPAs) offer a complementary principle: predict the representation of a missing or future target instead of reconstructing its raw observation (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024). V-JEPA 2 applies it to video understanding and action-conditioned planning, but its state is a video-derived latent rather than an explicit 3D scene (Assran et al., 2025). This raises the question: *can a JEPA learn the dynamics of an explicit Gaussian world while retaining the abstraction benefits of latent prediction?*

We propose 4DGS-JEPA, a Gaussian-native JEPA for causal multi-horizon prediction. Given only a causal prefix, it encodes time-indexed 3D Gaussians into a hierarchical predictive state with scene-, motion-group-, and Gaussian-level tokens, and a horizon-conditioned transition Φ_ϕ maps this state back into the same space, supporting both direct prediction and recursive rollout. Its central principle is *temporal composition*: a future state reached directly should agree with one reached by composing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shorter chronological transitions,

$$\Phi_\phi(Z_t, \delta_1 + \delta_2) \approx \Phi_\phi\big(\Phi_\phi(Z_t, \delta_1), \delta_2\big). \tag{1}$$

Endpoint supervision anchors direct predictions to future target embeddings, path supervision anchors recursive predictions, and composition suppresses residual dependence on the route taken. A *selective* geometry decoder then grounds the latent dynamics in group-level $\mathrm{SE}(3)$ motion and Gaussian deformation without reconstructing complete future appearance, and a hybrid correspondence mechanism handles the fact that the storage order of a Gaussian set does not define identity. Unlike recent Gaussian world models that generate or reconstruct future scenes (Lu et al., 2025; Chai et al., 2025; Lee et al., 2026; Zhang et al., 2026; Li et al., 2026), 4DGS-JEPA makes joint-embedding prediction its principal objective, operates natively on Gaussian states, and makes consistency across temporal prediction paths an explicit learning principle. Our contributions are:

- A Gaussian-native hierarchical JEPA for causal multi-horizon prediction, with scene-, motion-group-, and Gaussian-level predictive states and a closed, horizon-conditioned transition supporting direct prediction and recursive rollout (section 3);
- A temporally compositional learning framework combining endpoint supervision, recurrent multi-horizon path supervision, and explicit direct–composed consistency, together with a zero-loss path-agreement result and a finite-error rollout bound (section 3.3);
- A selective geometry decoder and geometry-level composition objective that ground latent dynamics in group-level $\mathrm{SE}(3)$ motion and Gaussian-level deformation without requiring complete future appearance reconstruction (section 3.4);
- A hybrid correspondence mechanism that preserves reliable persistent canonical identities and applies residual optimal transport only to unmatched or ambiguous Gaussian mass under reordering or topology change (section 4);
- Controlled experiments evaluating latent and geometric temporal composition, hybrid correspondence, and held-out future-frame rendering from predicted Gaussian geometry without RGB reconstruction supervision (section 5).

2 Related Work

Dynamic Gaussian scene representations. 3DGS represents a static scene by anisotropic Gaussians $(\mu_i, \Sigma_i, \alpha_i, c_i)$ optimized through differentiable rasterization (Kerbl et al., 2023). 4D-GS adds spatiotemporal features and a deformation field (Wu et al., 2024), and D-NeRF and HyperNeRF use canonical-space deformation for non-rigid motion and topology change (Pumarola et al., 2020; Park et al., 2021). These methods fit a representation to an observed sequence; 4DGS-JEPA instead learns, from a causal prefix $\mathcal{G}_{\le t}$, a predictive state Z_t that forecasts $\widehat{Z}_{t+\Delta}$.

Gaussian world models. GWM reconstructs action-conditioned future Gaussians with a 3D VAE and a latent diffusion transformer (Lu et al., 2025); GAF jointly reconstructs, predicts, and infers actions with motion-aware Gaussians (Chai et al., 2025); MoGaF forecasts long-horizon 4D Gaussians with motion-aware grouping (Lee et al., 2026); GaussianDream uses current and future Gaussian decoding as auxiliary supervision for manipulation (Zhang et al., 2026); and the Structured 4D Latent Predictive Model forecasts future 3D structure in a format-agnostic latent space (Li et al., 2026). All of these rely substantially on future-state generation or reconstruction, and none explicitly constrains alternative temporal paths to agree (positioning table in section A).

Joint-embedding predictive learning. JEPAs predict target embeddings from context embeddings (LeCun, 2022), for masked images (Assran et al., 2023), video (Bardes et al., 2024), planning (Assran et al., 2025), and dense features (Mur-Labadia et al., 2026). Variational JEPA makes latent prediction probabilistic (Huang, 2026b), and bidirectional JEPA enforces two-way predictive consistency (Huang, 2026a). 4DGS-JEPA transfers joint-embedding prediction to explicit Gaussian states and adds temporal composition.

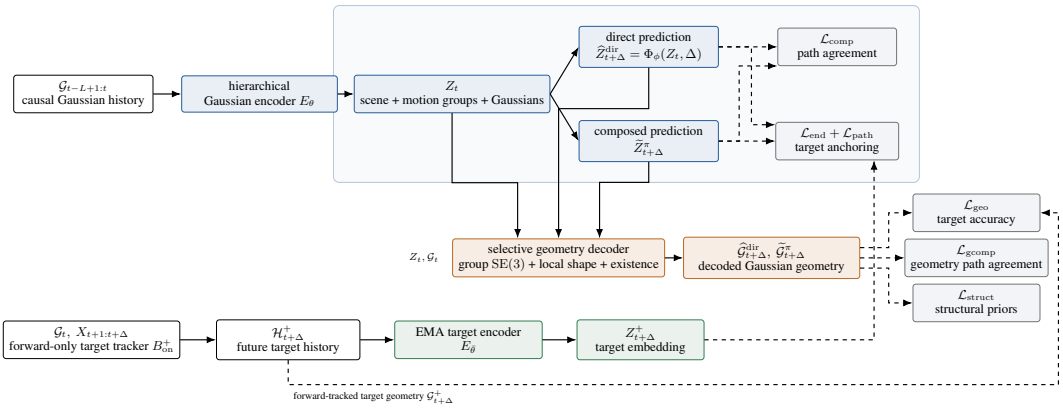


Figure 1: Overview of 4DGS-JEPA. A causal Gaussian history is encoded into a hierarchical predictive state with scene-, motion-group-, and Gaussian-level tokens. A shared horizon-conditioned transition produces direct and recursively composed future predictions. The future target history is constructed by forward-only tracking from the causal Gaussian state and subsequent observations, and an EMA target encoder provides the corresponding target embedding. Target embeddings provide endpoint and path supervision, while direct–composed agreement enforces temporal composition. A selective geometry decoder maps the predicted latent dynamics to explicit Gaussian motion, deformation, and existence variables. The decoded states are constrained by target-relative geometry accuracy, geometry-level composition, and source-relative structural priors. Solid arrows denote computational flow and dashed arrows denote supervision or consistency constraints.

3 Method

3.1 Causal Gaussian states and hierarchical predictive state

A dynamic scene at time t is a set of Gaussians

$$\mathcal{G}_t = \{g_{i,t}\}_{i=1}^{N_t}, \qquad g_{i,t} = (\mu_{i,t}, \Sigma_{i,t}, \alpha_{i,t}, c_{i,t}, f_{i,t}, e_{i,t}), \tag{2}$$

with center $\mu \in \mathbb{R}^3$, covariance $\Sigma \in \mathbb{S}^3_{++}$, opacity α, appearance c, optional semantic feature f, and existence $e \in [0,1]$. The prediction branch must be *causal*: context states $\mathcal{G}_{1:t}$ are built only from observations up to t, and target states after t come from a forward-only tracker that never revises states at or before t (section B). With context history $\mathcal{H}_t = (\mathcal{G}_{t-L+1}, \dots, \mathcal{G}_t)$ and a target history $\mathcal{H}^+_{t+\Delta}$ of the same length, a student encoder E_θ and an EMA teacher $E_{\bar\theta}$ ($\bar\theta \leftarrow \tau\bar\theta + (1-\tau)\theta$) define the predictive problem

$$\mathcal{H}_t \xrightarrow{E_\theta} Z_t \xrightarrow{\Phi_\phi(\cdot,\Delta)} \widehat{Z}_{t+\Delta} \approx Z^+_{t+\Delta} \xleftarrow{E_{\bar\theta}} \mathcal{H}^+_{t+\Delta}. \tag{3}$$

Future observations reach only the teacher, which is not needed at inference. The permutation-equivariant student represents the scene hierarchically,

$$Z_t = \big(z^S_t, \{z^O_{m,t}\}_{m=1}^{M_t}, \{z^G_{i,t}\}_{i\in\mathcal{I}_t}\big), \tag{4}$$

with a scene token, one token per motion group, and Gaussian tokens pooled bottom-up. The representation emphasizes geometry, opacity, existence, semantics, and low-order appearance; view-dependent appearance is excluded, so prediction does not reduce to reproducing appearance.

3.2 Horizon-conditioned transition and hierarchical distance

The transition $\widehat{Z}_{t+\Delta} = \Phi_\phi(Z_t, \Delta, A_{t:t+\Delta-1})$ (actions A optional) is *closed* in the predictive state space, so its output can be fed back for recursive rollout; Δ is embedded and conditions every block. States are compared by a hierarchical distance with group- and Gaussian-level correspondence plans Π^O, Π^G (section 4) of masses m_O, m_G:

$$\begin{aligned} D_{\mathrm{hier}}(Z, Z') = \lambda_S d_{\cos}(z^S, z'^S) + \tfrac{\lambda_O}{m_O+\varepsilon} \textstyle\sum_{mn} \Pi^O_{mn} d_{\cos}(z^O_m, z'^O_n) \\ + \tfrac{\lambda_G}{m_G+\varepsilon} \textstyle\sum_{ij} \Pi^G_{ij} \rho(z^G_i - z'^G_j), \end{aligned} \tag{5}$$

where $d_{\cos}$ is cosine distance and ρ a Huber penalty; the three terms capture global evolution, coherent group motion, and local deformation.

3.3 Endpoint, path, and temporal-composition objectives

Endpoint supervision anchors direct predictions for sampled horizons $\mathcal{K}_t$ with weights w_Δ:

$$\mathcal{L}_{\text{end}} = \textstyle\sum_{\Delta\in\mathcal{K}_t} w_\Delta \, D_{\text{hier}}\big(\widehat{Z}_{t+\Delta}, \text{stopgrad}(Z^+_{t+\Delta})\big). \tag{6}$$

Path supervision. Direct training does not ensure accuracy when Φ_ϕ is applied to its own outputs. For a partition $\pi = (\delta_1, \dots, \delta_J)$ of Δ with cumulative horizons $s_j = \sum_{q\le j}\delta_q$, the composed path $\widetilde{Z}^\pi_{t+s_j} = \Phi_\phi(\widetilde{Z}^\pi_{t+s_{j-1}}, \delta_j, A_{t+s_{j-1}:t+s_j-1})$, with $\widetilde{Z}^\pi_t = Z_t$, is anchored at every intermediate endpoint:

$$\mathcal{L}_{\text{path}} = \mathbb{E}_{\Delta,\pi}\Big[\textstyle\sum_j \omega_j \, D_{\text{hier}}\big(\widetilde{Z}^\pi_{t+s_j}, \text{stopgrad}(Z^+_{t+s_j})\big)\Big]. \tag{7}$$

Temporal composition. Target anchoring does not require two routes to the same time to agree. With the direct prediction $\widehat{Z}^{\text{dir}}_{t+s_j} = \Phi_\phi(Z_t, s_j, A_{t:t+s_j-1})$, we penalize disagreement in both stop-gradient directions,

$$\mathcal{L}_{\text{comp}} = \mathbb{E}_{\Delta,\pi}\Big[\textstyle\sum_j u_j \, D_{\text{hier}}\big(\widehat{Z}^{\text{dir}}_{t+s_j}, \text{stopgrad}(\widetilde{Z}^\pi_{t+s_j})\big) + \nu_j \, D_{\text{hier}}\big(\widetilde{Z}^\pi_{t+s_j}, \text{stopgrad}(\widehat{Z}^{\text{dir}}_{t+s_j})\big)\Big], \tag{8}$$

so that both routes receive consistency gradients (Huang, 2026a). For two intervals it encourages the controlled law $\Phi_\phi(Z_t, \delta_1+\delta_2, A_{t:t+\delta_1+\delta_2-1}) \approx \Phi_\phi(\Phi_\phi(Z_t, \delta_1, A_{t:t+\delta_1-1}), \delta_2, A_{t+\delta_1:t+\delta_1+\delta_2-1})$. We call this *temporal composition* rather than a semigroup property because transitions may depend on actions and absolute time and may include discrete events such as Gaussian birth or death. The three objectives are complementary: $\mathcal{L}_{\text{end}}$ and $\mathcal{L}_{\text{path}}$ anchor predictions to targets, whereas $\mathcal{L}_{\text{comp}}$ removes path dependence. Composition alone is insufficient, because two routes can agree while both are wrong or collapsed. This is made precise below; equality is on the supervised matched support, with normalized scene- and group-level embeddings (proofs in section D).

Proposition 1 (Zero-loss path agreement)**.** *Fix a cumulative horizon s_j and a partition π with $\omega_j > 0$ and $u_j + \nu_j > 0$. If the local contributions of s_j to $\mathcal{L}_{\text{path}}$ and $\mathcal{L}_{\text{comp}}$ are zero, then $\widehat{Z}^{\text{dir}}_{t+s_j} = \widetilde{Z}^\pi_{t+s_j} = Z^+_{t+s_j}$.*

Away from zero loss, recursion accumulates error. If segment j is locally L_j-Lipschitz with local error ϵ_j along a reference trajectory whose future states are the targets, and e_0 is the initial mismatch, then

$$\big\|\widetilde{Z}^\pi_{t+s_J} - Z^+_{t+s_J}\big\| \le \Big(\textstyle\prod_{q=1}^J L_q\Big)e_0 + \sum_{j=1}^J \Big(\prod_{q=j+1}^J L_q\Big)\epsilon_j, \tag{9}$$

so early local errors are amplified by the Lipschitz factors of all later transitions. Path supervision controls target-relative errors along realized rollouts, and composition constrains disagreement between routes; neither alone guarantees accurate long-horizon prediction.

3.4 Selective geometry decoder and geometry-level composition

Latent agreement does not guarantee coherent 3D motion. A decoder Q_ψ therefore translates a latent transition over $[a, b]$ into explicit Gaussian updates, predicting only the variables needed for geometric evolution (hence *selective*): per motion group, a rigid increment $\widehat{\xi}_{m,a\to b} \in \mathfrak{se}(3)$ with $\widehat{T}_{m,a\to b} = \text{Exp}(\widehat{\xi}_{m,a\to b}) = (\widehat{R}_m, \widehat{d}_m) \in \text{SE}(3)$; and per Gaussian, a residual $\widehat{r}_i \in \mathbb{R}^3$, a symmetric log-deformation $\widehat{S}_i$ with $\widehat{U}_i = \exp(\widehat{S}_i)$, and an existence probability $\widehat{e}_{i,b}$:

$$\widehat{\mu}_{i,b} = \widehat{R}_{m(i)}\mu_{i,a} + \widehat{d}_{m(i)} + \widehat{r}_i, \qquad \widehat{\Sigma}_{i,b} = \widehat{R}_{m(i)}\widehat{U}_i\Sigma_{i,a}\widehat{U}_i^\top\widehat{R}_{m(i)}^\top. \tag{10}$$

The congruence form keeps $\widehat{\Sigma}$ positive definite while allowing anisotropic deformation. Opacity, appearance, and semantics are transported, not reconstructed. Direct decoding applies Q_ψ once over $[t, t+s_j]$; composed decoding rolls the decoded Gaussians forward segment by segment. The *target-anchored geometry loss* $\mathcal{L}_{\text{geo}}$ compares each route's decoded state with the forward-tracked target through the correspondence plan, using Huber center errors, an SPD covariance distance d_{SPD}, and binary cross-entropy on existence (a dustbin supervises birth and death; section B). **Geometry-level**

composition. Two decoded routes can each be close to the target yet disagree with each other, and the same endpoint geometry can come from different splits between group motion and residuals. We therefore compose the decoded segment transforms, $\widehat{T}^{\pi}_{m,t\to t+\Delta} = \widehat{T}^{\pi}_{m,t+s_{J-1}\to t+s_J} \cdots \widehat{T}^{\pi}_{m,t\to t+s_1}$, and require agreement with the direct transform and geometry:

$$\mathcal{L}_{\mathrm{gcomp}} = \mathbb{E}_{\Delta,\pi}\Big[\textstyle\sum_m \big\|\mathrm{Log}\big((\widehat{T}^{\mathrm{dir}}_m)^{-1}\widehat{T}^{\pi}_m\big)\big\|_2 + \lambda_{\mathrm{gc}}\, D_{\mathrm{match}}\big(\widehat{\mathcal{G}}^{\mathrm{dir}}_{t+\Delta}, \widetilde{\mathcal{G}}^{\pi}_{t+\Delta}; \Pi^{G,\mathrm{dc}}\big)\Big], \tag{11}$$

where D_{match} sums matched center, covariance, and squared existence differences between the two routes. This is the geometric counterpart of $\mathcal{L}_{\mathrm{comp}}$. **Structural priors** $\mathcal{L}_{\mathrm{struct}}$ additionally preserve pairwise distances in groups identified as rigid, smooth non-rigid residuals between compatible neighbours, and penalize drift of static regions, all determined from the causal prefix only. **Representation regularization** $\mathcal{L}_{\mathrm{rep}}$ applies VICReg-style variance and covariance terms at each hierarchy level (Bardes et al., 2022). The full objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{end}} + \lambda_{\mathrm{path}}\mathcal{L}_{\mathrm{path}} + \lambda_{\mathrm{comp}}\mathcal{L}_{\mathrm{comp}} + \lambda_{\mathrm{geo}}\mathcal{L}_{\mathrm{geo}} + \lambda_{\mathrm{gcomp}}\mathcal{L}_{\mathrm{gcomp}} + \mathcal{L}_{\mathrm{struct}} + \lambda_{\mathrm{rep}}\mathcal{L}_{\mathrm{rep}}, \tag{12}$$

with no pixel-rendering loss and no complete-appearance reconstruction (training procedure in section E).

4 Gaussian Identity and Hybrid Correspondence

The losses above compare Gaussian sets, but a set's storage order does not define identity: $[A, B, C] \to [C, A, B]$ is the same scene, yet index matching compares A with C. Correspondence can also be one-to-many (split), many-to-one (merge), one-to-zero (death), or zero-to-one (birth). We therefore follow the principle

$$\text{reliable canonical identity} \Rightarrow \text{fixed correspondence}, \quad \text{uncertain identity} \Rightarrow \text{residual soft matching}. \tag{13}$$

A *prefix-anchored canonical bank* of N slots (active and reserved inactive) is built causally and tracked forward, so that ordinary evolution changes a slot's geometry and existence rather than its identifier. Valid persistent pairs $\mathcal{P}_{\mathrm{pers}}$ (same identity, still unique) define a partial plan $\Pi^{G,\mathrm{pers}}$, and only the residual marginals $\mathbf{p}_{\mathrm{res}} = \mathbf{p} - \Pi^{G,\mathrm{pers}}\mathbf{1}$ and $\mathbf{q}_{\mathrm{res}} = \mathbf{q} - (\Pi^{G,\mathrm{pers}})^{\top}\mathbf{1}$ enter entropic OT (Cuturi, 2013):

$$\Pi^{G,\mathrm{res}} = \arg\min_{\Pi \in \mathcal{U}(\mathbf{p}_{\mathrm{res}}, \mathbf{q}_{\mathrm{res}})} \langle \Pi, C\rangle + \varepsilon_{\mathrm{OT}} \textstyle\sum_{ij} \Pi_{ij}(\log \Pi_{ij} - 1), \tag{14}$$

so that $\Pi^G = \Pi^{G,\mathrm{pers}} + \Pi^{G,\mathrm{res}}$, with cost $C_{ij} = \beta_\mu \|\mu_i - \mu'_j\|^2 + \beta_f d_{\cos}(f_i, f'_j) + \beta_\Sigma d_{\mathrm{SPD}}(\Sigma_i, \Sigma'_j) + \beta_O(1 - \kappa^O_{m(i)m'(j)})$ that includes soft motion-group compatibility κ^O. Dustbins absorb unmatched mass, and fractional assignments accommodate splits and merges without imposing a wrong one-to-one match. Plans are built from Gaussian states and slots, never from the latent distance they are used to evaluate, and are treated as fixed assignments inside the losses, so the model cannot lower a loss by altering the correspondence (details in section C).

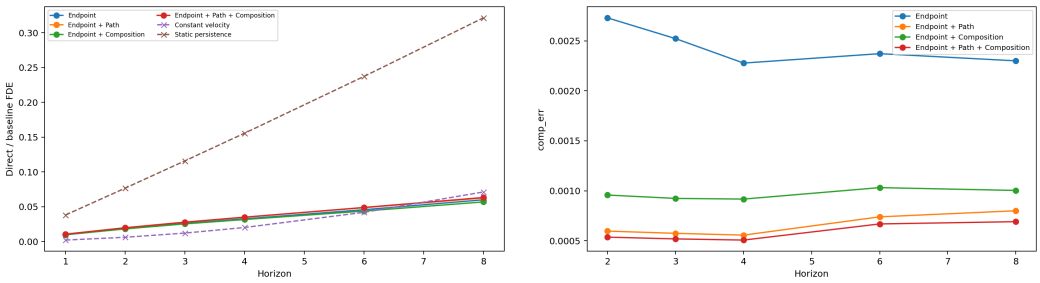
5 Experiments

We test four mechanism-level questions: (i) do path supervision and temporal composition reduce rollout drift and path dependence; (ii) does latent consistency transfer to decoded geometry; (iii) is hybrid correspondence reliable under reordering and topology change; and (iv) can predicted geometry be rendered into held-out future frames without RGB training?

Controlled dynamic Gaussian worlds. We procedurally generate trajectories of anisotropic 3D Gaussians in which motion groups move by rigid transformations with sampled velocities and accelerations, plus smooth Gaussian-specific deformation, $\mu_{i,t} = R_{m(i),t}\mu_{i,0} + d_{m(i),t} + r_{i,t}$ and $\Sigma_{i,t} = R_{m(i),t}U_{i,t}\Sigma_{i,0}U_{i,t}^{\top}R_{m(i),t}^{\top}$, with some static groups. This provides exact future geometry and identities without the confound of fitting a reconstruction model; train, validation, and test trajectories are sampled independently. Experiments 1–2 use context length $L = 4$, training horizons $\{1, 2, 4, 8\}$, unseen horizons $\{3, 6\}$, and fixed identities (existence fixed to one). Small attention-based models are used; comparable variants share capacity, data, optimizer, schedule, partitions, and three seeds

Table 1: Experiment 1 (3 seeds). Direct/rollout FDE and Δ_{roll} average over supervised horizons $\{1, 2, 4, 8\}$; composition metrics over non-trivial supervised paths; unseen FDE is rollout FDE at $\{3, 6\}$. Lower is better.

Method	Direct FDE	Rollout FDE	Δ_{roll}	CompErr	GeoCompErr	Unseen FDE
Endpoint	0.03013±0.00042	0.03061±0.00065	0.00048±0.00023	0.00244±0.00055	0.00739±0.00082	0.03621±0.00080
Endpoint + Path	0.03195±0.00243	0.03228±0.00256	0.00033±0.00014	0.00065±0.00007	**0.00649±0.00069**	0.03838±0.00304
Endpoint + Composition	**0.02905±0.00095**	**0.02940±0.00101**	0.00035±0.00014	0.00096±0.00011	0.00667±0.00074	**0.03466±0.00122**
Endpoint + Path + Comp.	0.03206±0.00245	0.03236±0.00255	**0.00029±0.00013**	**0.00058±0.00003**	0.00653±0.00089	0.03837±0.00283



(a) Direct FDE vs. horizon

(b) Direct–composed latent discrepancy

Figure 2: Experiment 1 across horizons. (a) Learned models strongly outperform static persistence, and Endpoint + Composition beats constant-velocity extrapolation at the longest horizon. (b) Endpoint supervision alone leaves substantial path dependence; path supervision and composition reduce it, and their combination gives the lowest CompErr.

(section F). We also report two causal analytic baselines: *static persistence* and *constant velocity*. **Metrics:** final displacement error (FDE) of direct and rollout predictions, the rollout gap $\Delta_{\text{roll}} = \text{FDE}_{\text{roll}} - \text{FDE}_{\text{dir}}$, and the latent and geometric path discrepancies $\text{CompErr} = D_{\text{hier}}(\widehat{Z}^{\text{dir}}_{t+\Delta}, \widetilde{Z}^{\pi}_{t+\Delta})$ and $\text{GeoCompErr} = \frac{1}{|\mathcal{I}|}\sum_i \|\widehat{\mu}^{\text{dir}}_i - \widetilde{\mu}^{\pi}_i\|_2$, computed on non-trivial paths.

5.1 Experiment 1: Path Supervision and Temporal Composition

We ablate the temporal latent losses (Endpoint; +Path; +Composition; +Path+Composition) with everything else fixed, including direct-route geometry supervision; $\mathcal{L}_{\text{gcomp}}$ is off (table 1 and fig. 2).

Composition improves consistency without sacrificing accuracy. Relative to Endpoint, Endpoint + Composition reduces direct and rollout FDE by 3.6% and 4.0% and CompErr by 60.6% (0.00244→0.00096). This is not a near-static solution: its direct FDE (0.02905) is far below static persistence (0.14802). Constant velocity is stronger on average on this smooth benchmark (0.02484), especially at short horizons, but at $\Delta = 8$ the learned model is better (0.05687 vs. 0.07120), where nonlinear motion accumulates.

Path supervision reduces rollout drift. Endpoint + Path lowers CompErr to 0.00065 without optimizing it directly, since both routes are anchored to common targets, and reduces Δ_{roll} by about 31%, at slightly higher FDE under the present weighting. Combining both gives the smallest rollout gap and CompErr (39% and 76% below Endpoint). Endpoint + Composition therefore gives the best accuracy–consistency trade-off, and the full objective gives the strongest consistency.

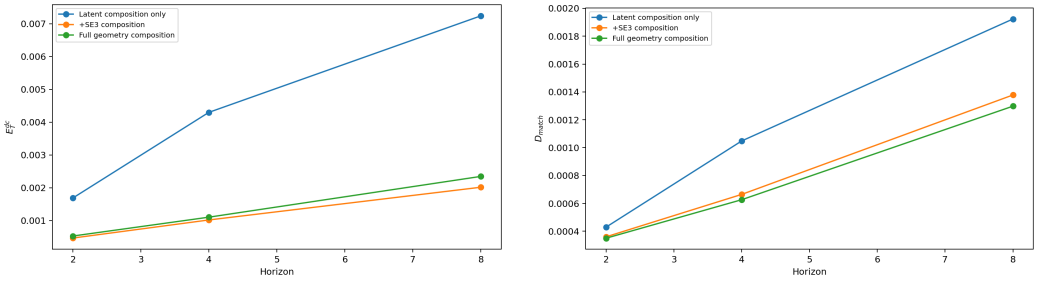
Generalization to unseen decompositions. On held-out partitions of seen segment lengths, the full model reduces CompErr from 0.01134 to 0.00260 (GeoCompErr 0.01428→0.01096); with unseen segment lengths (e.g., $(3, 5)$), CompErr drops from 0.00235 to 0.00057.

5.2 Experiment 2: Does Latent Composition Transfer to Gaussian Geometry?

Starting from the Endpoint + Composition configuration, with both routes geometry-supervised, we add the SE(3) term and then the full $\mathcal{L}_{\text{gcomp}}$. Using the true group motions, we measure the target error E_T^{target} and the direct–composed disagreement E_T^{dc} of group transforms (with a rotation–translation distance), together with FDE, covariance error, and D_{match} (table 2).

Table 2: Experiment 2 at $\Delta = 8$ (3 seeds). FDE, covariance error, and E_T^{target} are target-relative; E_T^{dc} and D_{match} measure direct–composed disagreement. Lower is better.

Method	Direct FDE	Direct cov. err.	E_T^{target}	E_T^{dc}	D_{match}
Latent composition only	**0.06614±0.00271**	0.19897±0.00387	0.26124±0.01038	0.00724±0.00057	0.00192±0.00015
+SE(3) composition	0.06675±0.00408	**0.19758±0.00330**	**0.25554±0.00119**	**0.00202±0.00031**	0.00138±0.00036
Full geometry composition	0.06728±0.00388	0.19758±0.00290	0.25582±0.00130	0.00235±0.00062	**0.00130±0.00012**



(a) Group-transform path discrepancy.

(b) Gaussian-level path discrepancy.

Figure 3: Explicit geometry consistency across non-trivial horizons. Panel (a) reports the direct–composed group-transform error E_T^{dc} in eq. (67); explicit $\mathrm{SE}(3)$ composition substantially reduces the discrepancy at every evaluated horizon. Panel (b) reports the Gaussian-level discrepancy D_{match} in eq. (42); the full geometry-composition objective gives the strongest primitive-level consistency. Explicit geometry regularization therefore further reduces residual path dependence beyond latent composition and common target anchoring.

Results. Latent composition alone does not fix a path-consistent motion decomposition: adding $\mathrm{SE}(3)$ composition cuts E_T^{dc} by 72.2% (0.00724→0.00202), and the full objective cuts the Gaussian-level D_{match} by 32.5%. Group-level composition thus aligns coarse structured motion, while Gaussian-level matching further aligns primitive geometry (a further 5.8% in D_{match}), consistently across horizons (fig. 3). The agreement is not due to trivial motion: all variants reach FDE 0.066–0.067 at $\Delta = 8$, versus 0.32158 for static persistence and 0.07120 for constant velocity, and $E_T^{\text{target}} \approx 0.256$ versus 0.34170 for identity transforms. Consistency gains cost at most 1.7% in FDE and slightly improve E_T^{target} (≈2%).

5.3 Experiment 3: Hybrid correspondence under identity perturbations

Independently of the learned transition, we perturb Gaussian sets with known lineage by permutation, splits, merges, deaths, activation of reserved slots, and their mixtures, and compare index matching, OT only, and hybrid correspondence. We report *LineageMass*, the fraction of real-to-real transport mass placed on ground-truth lineage edges, together with birth/death F1 and the geometry discrepancy induced by the plan (table 3).

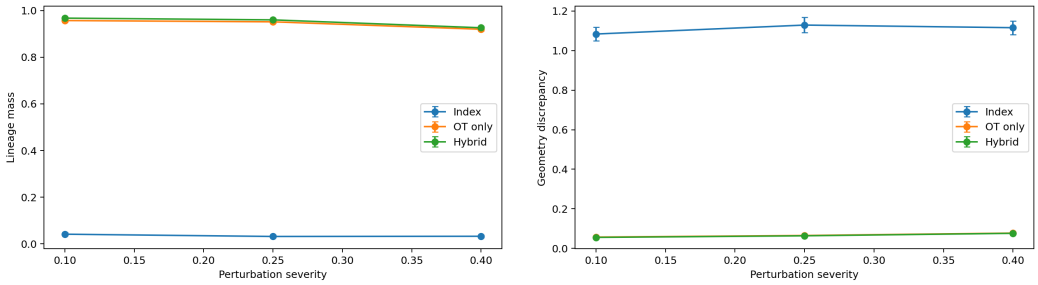
Results. Index matching fails even under pure permutation (LineageMass 0.041), confirming that array position is not identity. OT is a strong identity-agnostic baseline; Hybrid further raises split/merge LineageMass from 0.981 to 0.987 and lowers geometry discrepancy from 0.0501 to 0.0491, and it recovers permutations exactly. Under mixed perturbations, Hybrid beats OT at every severity (fig. 4), with a margin that narrows as more identities become invalid. When valid identities are deliberately hidden (0 to 50%), Hybrid degrades smoothly towards OT (LineageMass 0.960→0.956, versus 0.951 for OT), as intended by eq. (13) (section F.4).

5.4 Experiment 4: held-out future-frame rendering

We give each Gaussian a fixed RGB colour and opacity, generate 24-frame trajectories, and exclude the terminal frame I_T from all training and validation targets. Models are trained on horizons $\{1, 2, 4\}$ (with non-trivial partitions) and predict $\mathcal{G}_{T-L:T-1} \to Z_{T-1} \to \widehat{Z}_T \to \widehat{\mathcal{G}}_T \to \widehat{I}_T = \mathcal{R}(\widehat{\mathcal{G}}_T; \mathcal{C})$, where $\mathcal{R}$ is a standard Gaussian rasterizer (Kerbl et al., 2023) with a fixed camera. Appearance is transported,

Table 3: Gaussian-identity stress test (3 repetitions). LineageMass and F1: higher is better; geometry discrepancy: lower is better.

Method	Permutation lineage	Split/merge lineage	Birth/death F1	Geometry discrepancy
Index matching	0.04056 ± 0.00507	0.15563 ± 0.00907	0.61134 ± 0.00381	0.74311 ± 0.00449
OT only	0.99789 ± 0.00024	0.98100 ± 0.00099	$\mathbf{1.00000 \pm 0.00000}$	0.05010 ± 0.00014
Hybrid	$\mathbf{1.00000 \pm 0.00000}$	$\mathbf{0.98715 \pm 0.00103}$	$\mathbf{1.00000 \pm 0.00000}$	$\mathbf{0.04906 \pm 0.00013}$



(a) LineageMass under mixed perturbations.

(b) Induced geometry discrepancy.

Figure 4: Mixed correspondence stress test as perturbation severity increases. Hybrid maintains higher LineageMass and lower induced geometry discrepancy than OT only at every evaluated severity. As more persistent canonical identities become invalid, the margin decreases, consistent with Hybrid increasingly relying on residual OT.

so the test isolates whether *predicted geometry* projects to the correct image; no image-space loss is ever used. We also compare a direct $\Delta = 4$ prediction of the same endpoint with the composed route $(2, 2)$.

Results. Among the learned variants, Endpoint + Composition gives the best one-step accuracy and rendering (FDE 0.00946, PSNR 48.24 dB, SSIM 0.99901), with 76% lower FDE than last-frame persistence, so predicted geometry yields a close held-out observation without image supervision (table 4 and fig. 5). Constant velocity is much stronger on this smooth one-step task (PSNR 60.48 dB), so we read this as evidence of the *renderability* of learned predictive geometry, not of superior short-horizon video prediction. Composition remains visible at the rendered endpoint: CompErr falls by 76.4% and GeoCompErr by 19.0% relative to Endpoint, and the full model gives the strongest agreement.

6 Discussion and Limitations

Target anchoring vs. composition. The theory and experiments agree: $\mathcal{L}_{\text{comp}}$ constrains *how consistently* a future is reached, whereas target anchoring through $\mathcal{L}_{\text{end}}$ and $\mathcal{L}_{\text{path}}$ decides *whether it is correct* (proposition 1); their combination gives the smallest rollout gap and discrepancy. Latent consistency, moreover, does not automatically give consistent decoded motion, which explicit $\mathrm{SE}(3)$ and Gaussian-level composition supply.

Limitations. The current evidence is mechanism-level. Reconstruction error, large topology changes, and multimodal futures remain open, and composition constrains consistency rather than physical correctness or calibrated uncertainty (section G).

7 Conclusion

4DGS-JEPA takes a predictive rather than purely reconstructive view of dynamic Gaussian worlds: it learns how a Gaussian scene evolves, and requires one long transition and a sequence of shorter ones to the same future to yield compatible predictive states. A hierarchical Gaussian-native representation, a selective geometry decoder, geometry-level composition, and hybrid correspondence ground this principle in explicit 3D motion. Controlled experiments show reduced path dependence at

Table 4: Experiment 4: held-out terminal frame (learned: 3 seeds over 20 test trajectories). The last two columns compare direct and composed $(2, 2)$ routes to the same held-out endpoint. Bold: best overall; underline: best learned when the analytic baseline is stronger.

Method	Next FDE ↓	PSNR ↑	SSIM ↑	Δ=4 CompErr ↓	Δ=4 GeoCompErr ↓
Last-frame	0.03972±0.01203	35.30±3.24	0.98460±0.00849	—	—
Constant velocity	**0.00205±0.00074**	**60.48±3.21**	**0.99994±0.00005**	—	—
Endpoint	0.00972±0.00080	48.01±0.69	0.99896±0.00016	0.00867±0.00308	0.00802±0.00086
Endpoint + Path	0.01009±0.00089	47.74±0.95	0.99889±0.00021	0.00140±0.00020	0.00600±0.00105
Endpoint + Composition	0.00946±0.00079	48.24±0.84	0.99901±0.00017	0.00205±0.00047	0.00649±0.00133
Endpoint + Path + Comp.	0.01009±0.00085	47.72±0.95	0.99889±0.00021	**0.00126±0.00019**	**0.00583±0.00105**

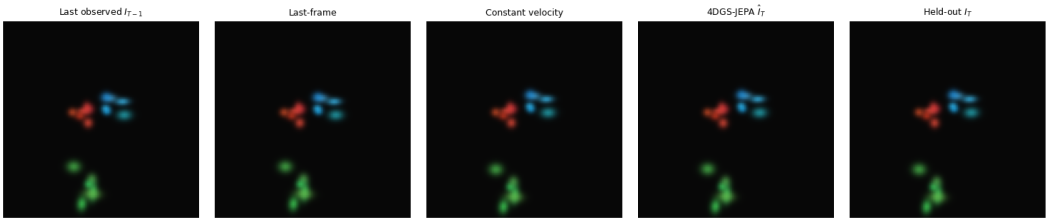


Figure 5: Held-out future-frame rendering (Endpoint + Composition, pre-specified example). Left to right: last observed frame I_{T-1}, last-frame persistence, constant velocity, the 4DGS-JEPA prediction $\widehat{I}_T$ obtained by rendering predicted Gaussian geometry, and the held-out target I_T. No RGB objective is used in training.

preserved accuracy, path-consistent decoded motion, robust correspondence under topology change, and renderable future geometry without pixel supervision. Scaling to causally reconstructed real scenes, learned appearance, multimodal futures, and action-conditioned prediction is the natural next step.

AI USE STATEMENT

AI tools were used to aid formatting and polish the writing of this paper, and to assist in developing and checking the proofs of the mathematical claims, namely the zero-loss path-agreement result (proposition 1) and the finite-error rollout bound (eq. (9)), whose complete derivations are given in section D. All proofs were verified by the authors. We have reviewed all AI-assisted content and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies predictive representation learning for dynamic 3D Gaussian scenes using procedurally generated synthetic data only; no human subjects, personal data, or real-world recordings are involved. We are not aware of risks specific to this work beyond those of general-purpose 3D scene modeling and world models.

REPRODUCIBILITY STATEMENT

Complete proofs of the theoretical results are given in section D, and the training procedure in section E. Section F specifies the procedural data generation, model architectures, optimization, regularization, and evaluation protocols for all four experiments, including seeds and per-seed results. Code will be released upon publication.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding

predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2. Featured Certification.

Ying Chai, Litao Deng, Ruizhi Shao, Jiajun Zhang, Kangchen Lv, Liangjun Xing, Xiang Li, Hongwen Zhang, and Yebin Liu. Gaf: Gaussian action field as a 4d representation for dynamic world modeling in robotic manipulation, 2025. URL https://arxiv.org/abs/2506.14135.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc.

Katlyn Edwards, Corrie Scalisi, Julianne DeMars-Smith, and Key Lee. Google colab for teaching cs and ml. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, SIGCSE 2024, pp. 1925, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704246. doi: 10.1145/3626253.3635432. URL https://doi.org/10.1145/3626253.3635432.

Yongchao Huang. Bijepa: Bi-directional joint embedding predictive architecture for symmetric representation learning, 2026a. URL https://arxiv.org/abs/2603.00049.

Yongchao Huang. VJEPA: Variational joint embedding predictive architectures as probabilistic world models. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=omJqMJt2fC.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), July 2023. ISSN 0730-0301. doi: 10.1145/3592433. URL https://doi.org/10.1145/3592433.

Yann LeCun. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Junmyeong Lee, Hoseung Choi, and Minsu Cho. Space-time forecasting of dynamic scenes with motion-aware gaussian grouping, 2026. URL https://arxiv.org/abs/2602.21668.

Zhiyi Li, Peilin Wu, Xiaoshen Han, Ruojin Cai, and Yilun Du. Structured 4d latent predictive model for robot planning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=aXAgpGfHGc.

Guanxing Lu, Baoxiong Jia, Puhao Li, Yixin Chen, Ziwei Wang, Yansong Tang, and Siyuan Huang. Gwm: Towards scalable gaussian world models for robotic manipulation, 2025. URL https://arxiv.org/abs/2508.17600.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning, 2026. URL https://arxiv.org/abs/2603.14482.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), December 2021. ISSN 0730-0301. doi: 10.1145/3478513.3480487. URL https://doi.org/10.1145/3478513.3480487.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes, 2020. URL https://arxiv.org/abs/2011.13961.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320, 2024. doi: 10.1109/CVPR52733.2024.01920.

Zijian Zhang, Yuqing Jiang, Qian Cheng, Xiaofan Li, Si Liu, Ding Zhao, Ping Luo, Weitao Zhou, and Haibao Yu. Gaussiandream: A feed-forward 3d gaussian world model for robotic manipulation, 2026. URL https://arxiv.org/abs/2605.20752.

A Extended Related Work and Positioning

Table 5 summarizes representative dynamic Gaussian scene representations, Gaussian world models, and joint-embedding predictive models. The key distinction of 4DGS-JEPA is that it combines a Gaussian-native predictive state with joint-embedding supervision and explicit temporal path consistency, without making complete future-scene reconstruction the principal learning objective.

A.1 Dynamic Gaussian scene representations

3D Gaussian Splatting (3DGS) represents a static scene using a set of anisotropic Gaussian primitives,

$$\mathcal{G} = \{g_i\}_{i=1}^{N}, \qquad g_i = (\mu_i, \Sigma_i, \alpha_i, c_i), \tag{15}$$

where μ_i and Σ_i denote the center and covariance of Gaussian i, while α_i and c_i denote its opacity and spherical-harmonic appearance coefficients. These parameters are optimized through differentiable rasterization for high-quality novel-view synthesis (Kerbl et al., 2023). The resulting representation describes a static scene but does not itself specify how that scene evolves over time.

Dynamic Gaussian representations extend this formulation to a time-indexed scene,

$$\mathcal{G}_t = \{g_{i,t}\}_{i=1}^{N_t},$$

in which Gaussian positions, covariances, opacities, appearances, or existence states may evolve. For example, 4D-GS augments 3DGS with spatiotemporal features and a deformation predictor for reconstruction and rendering at different timestamps (Wu et al., 2024). Related dynamic neural scene representations, including D-NeRF and HyperNeRF, similarly use canonical-space deformation to model non-rigid motion and topology change (Pumarola et al., 2020; Park et al., 2021).

Many such methods primarily optimize reconstruction or interpolation over an observed sequence: time indexes a scene representation fitted using observations from that sequence. 4DGS-JEPA addresses a different problem. Given only a causal Gaussian scene prefix $\mathcal{G}_{\leq t}$, it learns a predictive state Z_t and forecasts its future representation,

$$\mathcal{G}_{\leq t} \longrightarrow Z_t \longrightarrow \widehat{Z}_{t+\Delta}, \tag{16}$$

without requiring reconstruction of every future pixel or Gaussian attribute. Thus, 3DGS represents a static scene, dynamic Gaussian methods represent time-varying scenes, and 4DGS-JEPA focuses on learning reusable dynamics for causal future prediction.

A.2 Gaussian world models and forecasting

Recent Gaussian world models move beyond time-conditioned reconstruction towards explicit future prediction. GWM reconstructs action-conditioned future Gaussian states using a 3D VAE and

Table 5: Positioning relative to representative prior work. "Full reconstruction" indicates that decoding a complete future appearance or renderable state is central to the training objective. "Path comp." denotes explicit consistency between alternative temporal prediction paths.

Method	Native state	Future target	JEPA	Full recon.	Path comp.
4D-GS (Wu et al., 2024)	Gaussians	time query	no	yes	no
GWM (Lu et al., 2025)	Gaussians	future Gaussians	no	yes	no
GAF (Chai et al., 2025)	Gaussians	frame/action	no	yes	no
MoGaF (Lee et al., 2026)	Gaussians	future Gaussians	no	yes	no
Structured 4D (Li et al., 2026)	3D latent	future 3D	no	yes	no
V-JEPA 2 (Assran et al., 2025)	video latent	future latent	yes	no	no
4DGS-JEPA	Gaussian-native latent	latent/geometry	yes	no	yes

latent diffusion transformer and applies the learned model to imitation learning and model-based reinforcement learning (Lu et al., 2025). GAF augments Gaussians with motion attributes and jointly supports current-scene reconstruction, future prediction, and robot-action inference (Chai et al., 2025). MoGaF introduces motion-aware Gaussian grouping and forecasts long-horizon dynamic-scene evolution (Lee et al., 2026). GaussianDream trains a feed-forward world-model representation using current and future Gaussian decoding as supervision and discards the decoders at test time for manipulation (Zhang et al., 2026).

These approaches demonstrate the value of explicit 3D future prediction, but they rely substantially on explicit future-state generation or reconstruction supervision. 4DGS-JEPA instead makes prediction of future target embeddings its principal learning objective. Selective geometry decoding provides structured geometric grounding without requiring reconstruction of the complete future Gaussian appearance.

The Structured 4D Latent Predictive Model is particularly close in spirit: it forecasts future 3D scene structure in a general latent space conditioned on observations and language and decodes that representation into point clouds or Gaussians for planning (Li et al., 2026). However, its predictive representation is not Gaussian-native and it does not explicitly constrain alternative temporal prediction paths to agree. In contrast, 4DGS-JEPA learns a Gaussian-native predictive representation and explicitly regularizes both latent and decoded dynamics to be consistent across alternative chronological decompositions of the same future evolution.

A.3 Joint-embedding predictive learning

JEPA proposes learning representations by predicting compatible target embeddings from context embeddings rather than reconstructing observations (LeCun, 2022). I-JEPA instantiates this principle for masked image regions (Assran et al., 2023), while V-JEPA extends feature prediction to video without pixel reconstruction (Bardes et al., 2024). V-JEPA 2 combines large-scale video pretraining with an action-conditioned latent dynamics model for robotic planning (Assran et al., 2025), and V-JEPA 2.1 further emphasizes dense, spatially grounded predictive features (Mur-Labadia et al., 2026).

Variational JEPA (VJEPA) extends deterministic joint-embedding prediction to a predictive distribution over future latent states, enabling uncertainty-aware world modelling without requiring an observation-space likelihood (Huang, 2026b). 4DGS-JEPA is complementary to this probabilistic direction. Its focus is instead on transferring joint-embedding predictive learning to explicit Gaussian scene states and enforcing temporal composition across direct and recursively composed future predictions.

B Extended Method Details

B.1 Causal construction of context and target states

An important requirement is that the Gaussian state available to the predictor be constructed *causally*. For a context endpoint t, an online Gaussian front-end B_{on} may use only observations up to t:

$$\mathcal{G}_{1:t} = B_{\mathrm{on}}(X_{1:t}). \tag{17}$$

Starting from the Gaussian state available at t, a target-side tracker is rolled forward only with observations available up to each target endpoint,

$$\mathcal{G}^{+}_{t+1:t+\Delta} = B^{+}_{\mathrm{on}}(\mathcal{G}_t, X_{t+1:t+\Delta}), \tag{18}$$

and is not allowed to revise any state at or before t. A Gaussian representation fitted jointly using the complete sequence would generally violate this protocol because its representation at time t could depend on future observations. The controlled experiments instantiate this causal protocol directly in Gaussian space, where the generated state at each time is available without fitting a reconstruction model; evaluation with a learned online Gaussian front-end is left to future work.

For a context length L, define the causal context history

$$\mathcal{H}_t = (\mathcal{G}_{t-L+1}, \ldots, \mathcal{G}_t). \tag{19}$$

For every target horizon Δ, the target branch receives a history of the same temporal extent,

$$\mathcal{H}^{+}_{t+\Delta} = (\mathcal{G}^{+}_{t+\Delta-L+1}, \ldots, \mathcal{G}^{+}_{t+\Delta}), \tag{20}$$

where states at or before the context endpoint are inherited unchanged, $\mathcal{G}^{+}_{k} = \mathcal{G}_k$ for $k \le t$, while states after t are obtained through forward-only target tracking. The goal is to learn a predictive state that is sufficient for forecasting future scene evolution but need not preserve unpredictable or unnecessary appearance details (Huang, 2026b).

B.2 Encoder and transition details

where the superscripts S, O, and G denote the scene, motion-group, and Gaussian levels, respectively. Here, the scene token z^{S}_{t} summarizes the global predictive state, the motion-group token $z^{O}_{m,t}$ represents group m, and the Gaussian token $z^{G}_{i,t}$ represents primitive i. The quantity M_t is the number of motion groups at time t, while $\mathcal{I}_t \subseteq \{1, \ldots, N_t\}$ indexes the Gaussian primitives represented explicitly by local tokens and may contain all active primitives or a sampled subset for computational efficiency.

The encoder is permutation equivariant over Gaussian primitives. It may use canonical coordinates, motion-group membership, temporal information, and persistent identity embeddings when available. Pooling follows the same hierarchy: Gaussian-level tokens first aggregate into motion-group tokens, which subsequently aggregate into the global scene token. By default, the predictive representation emphasizes geometry, opacity, existence, semantics, and low-order appearance cues. Detailed view-dependent appearance is excluded or independently perturbed so that joint-embedding prediction does not reduce to reproducing exact appearance.

where $\tau \in [0, 1)$ is the EMA momentum and the superscript $+$ denotes target-branch quantities. The teacher is not updated directly by gradient descent and is not required for future prediction at inference time. We use equal context and target history lengths so that the student and teacher representations have access to comparable temporal information; for example, a multi-frame teacher can represent motion information that would be unavailable to a single-frame target encoder.

Given the hierarchical predictive state Z_t, we learn a horizon-conditioned transition operator

$$\widehat{Z}_{t+\Delta} = \Phi_{\phi}(Z_t, \Delta, A_{t:t+\Delta-1}), \tag{21}$$

where $A_{t:t+\Delta-1}$ is an optional action sequence and is omitted for passive scene forecasting. The transition is *closed* in the predictive state space: it consumes and returns the same hierarchical state type as eq. (4). Consequently, a predicted state can itself be supplied to Φ_{ϕ} as the input to a subsequent transition, enabling recursive rollout and temporal composition. The relative horizon Δ is mapped to a temporal embedding that conditions each transition block, specifying how far forward the operator should propagate the current state. When the dynamics are explicitly non-autonomous, an absolute-time embedding of the transition start time t is additionally provided.

B.3 HORIZON WEIGHTING, PATH ROLLOUT, AND COMPOSITION

For discrete frame horizons, a natural choice is the normalized temporal discount

$$w_\Delta = \frac{\gamma^{\Delta-1}}{\sum_{\delta\in\mathcal{K}_t}\gamma^{\delta-1}}, \qquad \gamma \in (0,1]. \tag{22}$$

Here, $\gamma = 1$ weights all sampled horizons equally, whereas $\gamma < 1$ progressively emphasizes nearer and typically more predictable future states. The weighting in eq. (22) therefore provides a general mechanism for emphasizing different temporal ranges. In the controlled experiments below, however, we use one terminal horizon per training step, cycling uniformly through the supervised horizon set, and do not separately ablate γ. Unlike pixel- or Gaussian-reconstruction objectives, $\mathcal{L}_{\text{end}}$ requires the transition to predict only the target *representation* of the future Gaussian world.

Direct multi-horizon supervision does not guarantee that the learned transition remains accurate when recursively applied to its own predictions. For each terminal horizon Δ, we therefore sample a temporal partition

$$\pi = (\delta_1, \dots, \delta_J), \qquad \delta_j > 0, \qquad \sum_{j=1}^{J}\delta_j = \Delta. \tag{23}$$

Let

$$s_0 = 0, \qquad s_j = \sum_{q=1}^{j}\delta_q, \qquad s_J = \Delta \tag{24}$$

denote the corresponding cumulative horizons. Starting from $\widetilde{Z}^\pi_t = Z_t$, the recursively composed path is

$$\widetilde{Z}^\pi_{t+s_j} = \Phi_\phi\left(\widetilde{Z}^\pi_{t+s_{j-1}}, \delta_j, A_{t+s_{j-1}:t+s_j-1}\right). \tag{25}$$

where $\omega_j \geq 0$ controls the contribution of the jth cumulative horizon and is normalized over the sampled path.

Unlike direct endpoint prediction, the composed path repeatedly applies Φ_ϕ to its own predicted states. Supervising the intermediate states therefore exposes the transition to the state distribution encountered during rollout and reduces the mismatch between direct training and recursive inference.

For each cumulative horizon s_j, there are two predictions of the same future state. The *direct* prediction reaches $t + s_j$ in a single transition,

$$\widehat{Z}^{\text{dir}}_{t+s_j} = \Phi_\phi\big(Z_t, s_j, A_{t:t+s_j-1}\big), \tag{26}$$

whereas the *composed* prediction $\widetilde{Z}^\pi_{t+s_j}$ is obtained recursively through a sequence of shorter transitions along the temporal partition $\pi = (\delta_1, \dots, \delta_J)$ according to eq. (25). Thus, $\widehat{Z}^{\text{dir}}_{t+s_j}$ represents one direct transition to the future endpoint, while $\widetilde{Z}^\pi_{t+s_j}$ represents a composition of shorter transitions reaching the same endpoint. Temporal composition requires these two predictions to agree.

For two consecutive intervals, the objective encourages the controlled composition law

$$\Phi_\phi(Z_t, \delta_1+\delta_2, A_{t:t+\delta_1+\delta_2-1}) \approx \Phi_\phi(\Phi_\phi(Z_t, \delta_1, A_{t:t+\delta_1-1}), \delta_2, A_{t+\delta_1:t+\delta_1+\delta_2-1}). \tag{27}$$

We refer to this property as *temporal composition* rather than an exact one-parameter semigroup property because the transition may depend on control inputs and absolute time and may include discrete state changes such as Gaussian birth or death. The relevant requirement is therefore consistency across admissible decompositions of the same controlled temporal evolution. In passive forecasting, the action arguments are omitted.

B.4 SELECTIVE GEOMETRY DECODER

The latent objectives above define the predictive representation without requiring observation reconstruction. Latent agreement alone, however, does not guarantee that a predicted state corresponds

to coherent motion in 3D. We therefore introduce a *selective geometry decoder* [1] that grounds the latent transition in a small set of interpretable Gaussian updates. Importantly, the decoder does not reconstruct the complete future Gaussian appearance; it predicts only the variables required to describe geometric evolution.

For any temporal interval $[a, b]$ with $a < b$, the decoder receives the latent state at the start of the interval, Z_a, the predicted latent state at its endpoint, Z_b, the source Gaussian scene $\mathcal{G}_a$, and the elapsed time $b - a$. Its role is to translate the latent transition from a to b into explicit motion of the scene's motion groups and Gaussian primitives.

Let $\mathcal{I}_a^{\mathrm{dec}}$ denote the decoder-side canonical slots at time a, including represented active primitives and any reserved inactive slots eligible for activation. The shared decoder Q_ψ produces

$$\left(\{\widehat{\xi}_{m,a\to b}\}_{m=1}^{M_a}, \{\widehat{r}_{i,a\to b}, \widehat{S}_{i,a\to b}, \widehat{e}_{i,b}\}_{i\in\mathcal{I}_a^{\mathrm{dec}}}\right) = Q_\psi(Z_a, Z_b, \mathcal{G}_a, b-a), \tag{28}$$

where a and b denote the start and end times of the decoded interval, $m \in \{1, \dots, M_a\}$ indexes motion groups present at time a, and $i \in \mathcal{I}_a^{\mathrm{dec}}$ indexes decoder-side Gaussian slots. The notation $a \to b$ indicates that a quantity describes the change accumulated over the interval from a to b.

At the group level, $\widehat{\xi}_{m,a\to b} \in \mathfrak{se}(3)$ is the predicted rigid-motion increment for motion group m. At the Gaussian level, $\widehat{r}_{i,a\to b} \in \mathbb{R}^3$ is a local non-rigid displacement residual, $\widehat{S}_{i,a\to b} \in \mathbb{S}^3$ is a symmetric log-deformation controlling covariance change, and $\widehat{e}_{i,b} \in [0, 1]$ is the predicted existence probability of slot i at the endpoint b. Thus, the decoder decomposes scene evolution into coherent group motion, local primitive deformation, covariance evolution, and birth/death dynamics. During forecasting, Z_b is always a predicted endpoint state, either direct or composed; the teacher representation is not supplied to the decoder.

For a Gaussian i assigned to motion group $m(i)$, the predicted group-level motion increment $\widehat{\xi}_{m,a\to b} \in \mathfrak{se}(3)$ is converted into a finite rigid transformation through the exponential map,

$$\widehat{T}_{m,a\to b} = \mathrm{Exp}(\widehat{\xi}_{m,a\to b}) = \left(\widehat{R}_{m,a\to b}, \widehat{d}_{m,a\to b}\right) \in \mathrm{SE}(3), \tag{29}$$

where $\mathfrak{se}(3)$ is the Lie algebra of 3D rigid motions, $\widehat{R}_{m,a\to b} \in \mathrm{SO}(3)$ is the predicted rotation of group m from time a to b, and $\widehat{d}_{m,a\to b} \in \mathbb{R}^3$ is its predicted translation. Thus, $\mathrm{Exp} : \mathfrak{se}(3) \to \mathrm{SE}(3)$ maps the group-level motion increment to a rigid transformation composed of rotation and translation.

At the Gaussian level, the symmetric log-deformation is mapped through the matrix exponential to the local deformation matrix

$$\widehat{U}_{i,a\to b} = \exp(\widehat{S}_{i,a\to b}), \tag{30}$$

where $\widehat{U}_{i,a\to b} \in \mathbb{S}^3_{++}$ is a positive-definite local deformation, or stretch, matrix controlling Gaussian-specific changes in shape and scale. Thus, $\widehat{T}_{m,a\to b}$ captures the coherent rigid motion shared by a motion group, while $\widehat{U}_{i,a\to b}$ and $\widehat{r}_{i,a\to b}$ capture local deviations from that group motion. Here, $\mathrm{Exp}(\cdot)$ denotes the rigid-motion exponential map from $\mathfrak{se}(3)$ to $\mathrm{SE}(3)$, whereas $\exp(\cdot)$ in eq. (30) denotes the matrix exponential.

The corresponding geometric updates for the center and covariance of Gaussian i are

$$\widehat{\mu}_{i,b} = \widehat{R}_{m(i),a\to b}\mu_{i,a} + \widehat{d}_{m(i),a\to b} + \widehat{r}_{i,a\to b}, \tag{31}$$

$$\widehat{\Sigma}_{i,b} = \widehat{R}_{m(i),a\to b}\widehat{U}_{i,a\to b}\Sigma_{i,a}\widehat{U}_{i,a\to b}^\top\widehat{R}_{m(i),a\to b}^\top. \tag{32}$$

The center update therefore combines the shared group rotation and translation with a Gaussian-specific residual displacement. The covariance update first applies the local shape/scale deformation and then transports the resulting covariance through the group rotation. Its congruence form preserves positive definiteness while allowing contraction, expansion, and anisotropic deformation.

Together with the predicted existence probability $\widehat{e}_{i,b}$ and the transported source attributes, these updates instantiate the decoded Gaussian at time b,

$$\widehat{g}_{i,b} = \left(\widehat{\mu}_{i,b}, \widehat{\Sigma}_{i,b}, \alpha_{i,a}, c_{i,a}, f_{i,a}, \widehat{e}_{i,b}\right), \qquad \widehat{\mathcal{G}}_b = \{\widehat{g}_{i,b}\}_{i\in\mathcal{I}_a^{\mathrm{dec}}}, \tag{33}$$

[1] Here, *selective* means selecting which Gaussian attributes are decoded: the decoder predicts the variables needed for geometric dynamics rather than reconstructing the complete future Gaussian state or appearance. It doesn't mean the decoder is *optional*.

with inactive slots interpreted according to their predicted existence probabilities. Thus, the decoder outputs are converted into an explicit predicted Gaussian scene $\widehat{\mathcal{G}}_b$. This scene is subsequently used for target-anchored geometric supervision and geometry-level composition; for a composed rollout, it additionally becomes the source Gaussian state for decoding the next temporal segment.

The full decoding chain can be summarized as

$$(Z_a, Z_b, \mathcal{G}_a, b-a) \xrightarrow{Q_\psi} \Big(\{\widehat{\xi}_{m,a\to b}\}, \{\widehat{r}_{i,a\to b}, \widehat{S}_{i,a\to b}, \widehat{e}_{i,b}\}\Big) \xrightarrow{\text{geometric update}} \widehat{\mathcal{G}}_b \xrightarrow{\text{losses / next rollout step}} \cdots \tag{34}$$

That is, the decoder Q_ψ first predicts motion and deformation variables, these variables are then converted into an explicit Gaussian state $\widehat{\mathcal{G}}_b$ through eqs. (10) and (32), and the resulting state is subsequently used either for geometric supervision or as the source state for the next segment in a composed rollout.

The default decoder does not predict high-frequency appearance. Appearance coefficients $c_{i,a}$ and semantic features $f_{i,a}$ are transported from the source state where defined, while the predicted existence probability $\widehat{e}_{i,b}$ controls activation and deactivation of canonical slot i at the endpoint b. Opacity $\alpha_{i,a}$ is likewise transported for surviving primitives rather than independently reconstructed. A small appearance-residual head may be trained separately for rendering evaluation, but appearance reconstruction is excluded from the default JEPA objective.

For a cumulative horizon s_j, direct decoding applies the shared decoder once,

$$\widehat{\mathcal{G}}^{\mathrm{dir}}_{t+s_j} = Q^{\mathcal{G}}_\psi\Big(Z_t, \widehat{Z}^{\mathrm{dir}}_{t+s_j}, \mathcal{G}_t, s_j\Big), \tag{35}$$

where $Q^{\mathcal{G}}_\psi$ denotes application of the decoder outputs in eq. (28) through the geometric updates in eqs. (10) and (32). For a composed path, we initialize $\widetilde{\mathcal{G}}^{\pi}_t = \mathcal{G}_t$ and recursively decode each interval,

$$\widetilde{\mathcal{G}}^{\pi}_{t+s_q} = Q^{\mathcal{G}}_\psi\Big(\widetilde{Z}^{\pi}_{t+s_{q-1}}, \widetilde{Z}^{\pi}_{t+s_q}, \widetilde{\mathcal{G}}^{\pi}_{t+s_{q-1}}, \delta_q\Big), \qquad q = 1, \dots, j. \tag{36}$$

Thus, group transformations, local residuals, covariance changes, and existence states are rolled forward along the composed path rather than decoded only at the terminal endpoint.

B.5 Target-anchored geometric supervision

The selective decoder is trained against the forward-tracked target Gaussian states. The purpose of this supervision is to ask whether the decoded prediction places the correct Gaussian geometry at the future endpoint: predicted primitives should have the correct positions, shapes, and existence states.

Let $\mathcal{I}_{\mathrm{pred}}$ denote the predicted decoder slots under comparison and let $\mathcal{I}^+$ denote the real target primitives, excluding the dustbin. Because predicted and target Gaussians are not necessarily index-aligned, we compare them through the Gaussian transport plan Π^G, where Π^G_{ij} measures the correspondence mass between predicted slot i and target primitive j. Given a predicted Gaussian state $\widehat{\mathcal{G}}$, a target state $\mathcal{G}^+$, and their transport plan Π^G, we define

$$\ell_{\mathrm{geo}}\Big(\widehat{\mathcal{G}}, \mathcal{G}^+; \Pi^G\Big) = \sum_{i\in\mathcal{I}_{\mathrm{pred}}} \sum_{j\in\mathcal{I}^+} \Pi^G_{ij}\Big[\lambda_\mu \rho_\delta(\widehat{\mu}_i - \mu^+_j) + \lambda_\Sigma d_{\mathrm{SPD}}(\widehat{\Sigma}_i, \Sigma^+_j)\Big] + \lambda_e \sum_{i\in\mathcal{I}_{\mathrm{pred}}} \mathrm{BCE}\Big(\widehat{e}_i, \overline{e}^+_i\Big). \tag{37}$$

The loss measures three complementary geometric discrepancies. The center term $\rho_\delta(\widehat{\mu}_i - \mu^+_j)$ measures the positional error between predicted Gaussian i and target Gaussian j using the robust Huber penalty ρ_δ. The covariance term $d_{\mathrm{SPD}}(\widehat{\Sigma}_i, \Sigma^+_j)$ measures the discrepancy between their Gaussian covariance ellipsoids, and therefore penalizes errors in shape, scale, and orientation. Both terms are weighted by the correspondence mass Π^G_{ij}, so strongly matched Gaussian pairs contribute more strongly to the loss. The coefficients $\lambda_\mu, \lambda_\Sigma \geq 0$ control the relative strengths of center and covariance supervision.

The final term supervises whether each predicted Gaussian slot should exist at the target endpoint. Here, $\mathrm{BCE}(p, y) = -y \log p - (1-y)\log(1-p)$ denotes binary cross-entropy between the predicted existence probability $\widehat{e}_i$ and a soft target existence value $\overline{e}^+_i$. We define

$$\overline{e}^+_i = \frac{\sum_{j\in\mathcal{I}^+} \Pi^G_{ij}}{\sum_{j\in\mathcal{I}^+\cup\{\varnothing\}} \Pi^G_{ij} + \varepsilon_{\mathrm{match}}}, \tag{38}$$

where $\varnothing$ denotes the target dustbin and $\varepsilon_{\text{match}} > 0$ prevents division by zero. The numerator is the transport mass from predicted slot i assigned to real target geometry, while the denominator additionally includes mass assigned to the dustbin. Consequently, $\overline{e}_i^+ \approx 1$ indicates that slot i should remain or become active, whereas $\overline{e}_i^+ \approx 0$ indicates that it should disappear or remain inactive. Intermediate values naturally arise from soft or ambiguous correspondences. Transport to the dustbin therefore supervises disappearance or death, whereas transport from a reserved inactive slot to a real target primitive supervises birth and the geometry of the newly activated slot.

The target-anchored geometry objective applies this comparison to both prediction routes at every cumulative endpoint:

$$\mathcal{L}_{\text{geo}} = \mathbb{E}_{\Delta,\pi}\left[\sum_{j=1}^{J} \eta_j \left(\ell_{\text{geo}}\left(\widehat{\mathcal{G}}^{\text{dir}}_{t+s_j}, \mathcal{G}^+_{t+s_j}; \Pi^G_{j,\text{dir}}\right) + \ell_{\text{geo}}\left(\widetilde{\mathcal{G}}^{\pi}_{t+s_j}, \mathcal{G}^+_{t+s_j}; \Pi^G_{j,\pi}\right)\right)\right]. \tag{39}$$

Thus, at each cumulative horizon s_j, the direct prediction $\widehat{\mathcal{G}}^{\text{dir}}_{t+s_j}$ and the composed prediction $\widetilde{\mathcal{G}}^{\pi}_{t+s_j}$ are each compared independently with the same forward-tracked target geometry $\mathcal{G}^+_{t+s_j}$. The corresponding transport plans $\Pi^G_{j,\text{dir}}$ and $\Pi^G_{j,\pi}$ account for potentially different Gaussian correspondences along the two prediction routes. The weights $\eta_j \geq 0$ control the contribution of each cumulative endpoint and are normalized over the sampled path, while the expectation averages over sampled terminal horizons Δ and temporal partitions π.

Importantly, $\mathcal{L}_{\text{geo}}$ measures *target accuracy*: it asks whether each prediction route produces the correct future geometry. It does not directly compare the direct and composed routes with one another; their mutual geometric agreement is enforced separately by the geometry-level composition objective in section 3.4. We use the affine-invariant SPD distance $d_{\text{SPD}}(\Sigma_1, \Sigma_2) = \|\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})\|_F$ in the full implementation and a log-Cholesky surrogate for large batches. In this way, the latent objectives determine what future information should be predicted, while the selective decoder encourages those predictions to correspond to coherent evolution of explicit Gaussian geometry.

B.6 Geometry-level composition

The latent composition loss $\mathcal{L}_{\text{comp}}$ in section 3.3 requires direct and composed *latent* predictions to agree. This does not by itself guarantee that decoding the two latent paths yields the same explicit 3D motion. We therefore impose an analogous consistency constraint in Gaussian geometry. Whereas the target-anchored geometry loss $\mathcal{L}_{\text{geo}}$ asks whether each decoded path agrees with the future target geometry, the geometry-level composition loss introduced below asks whether the *direct and composed decoded motions agree with each other*.

Recall from eq. (29) that $\widehat{T}_{m,a\to b} \in \text{SE}(3)$ denotes the predicted rigid transformation of motion group m over an interval $[a, b]$. For a terminal horizon Δ and temporal partition $\pi = (\delta_1, \ldots, \delta_J)$ with cumulative horizons $s_J = \Delta$, let $\widehat{T}^{\text{dir}}_{m,t\to t+\Delta}$ denote the transformation obtained by decoding the direct prediction from t to $t + \Delta$. The corresponding composed transformation is obtained by multiplying the shorter segment transformations in chronological order,

$$\widehat{T}^{\pi}_{m,t\to t+\Delta} = \widehat{T}^{\pi}_{m,t+s_{J-1}\to t+s_J} \cdots \widehat{T}^{\pi}_{m,t+s_1\to t+s_2} \widehat{T}^{\pi}_{m,t\to t+s_1}. \tag{40}$$

The rightmost transformation is applied first. Thus, $\widehat{T}^{\text{dir}}_{m,t\to t+\Delta}$ represents one decoded rigid motion over the full horizon, whereas $\widehat{T}^{\pi}_{m,t\to t+\Delta}$ represents the same-horizon motion obtained by composing the shorter decoded transformations along path π.

To compare two rigid transformations, we use their relative transformation. For $T_1, T_2 \in \text{SE}(3)$, define

$$d_{\text{SE}(3)}(T_1, T_2) = \left\|\text{Log}\left(T_1^{-1}T_2\right)\right\|_2, \tag{41}$$

where $\text{Log} : \text{SE}(3) \to \mathfrak{se}(3)$ is the Lie-group logarithm. The relative transform $T_1^{-1}T_2$ describes the residual rigid motion required to move from T_1 to T_2, and $d_{\text{SE}(3)}(T_1, T_2) = 0$ when the transformations coincide. Thus, this distance measures disagreement in group-level rotation and translation.

We additionally compare the Gaussian-level geometry produced by the two paths. Let $\Pi^{G,\text{dc}}$ denote a direct–composed Gaussian correspondence plan, where $\Pi^{G,\text{dc}}_{ij}$ measures the correspondence mass

between Gaussian i in the direct prediction and Gaussian j in the composed prediction. Analogously to the target-anchored geometry loss in eq. (37), we define

$$D_{\mathrm{match}}\left(\widehat{\mathcal{G}}^{\mathrm{dir}}, \widetilde{\mathcal{G}}^{\pi}; \Pi^{G,\mathrm{dc}}\right) = \sum_{i,j} \Pi_{ij}^{G,\mathrm{dc}} \left[\lambda_\mu \rho_\delta\left(\widehat{\mu}_i^{\mathrm{dir}} - \widetilde{\mu}_j^{\pi}\right) + \lambda_\Sigma d_{\mathrm{SPD}}\left(\widehat{\Sigma}_i^{\mathrm{dir}}, \widetilde{\Sigma}_j^{\pi}\right) + \lambda_e \left(\widehat{e}_i^{\mathrm{dir}} - \widetilde{e}_j^{\pi}\right)^2\right]. \tag{42}$$

Here, $\widehat{\mu}_i^{\mathrm{dir}}$, $\widehat{\Sigma}_i^{\mathrm{dir}}$, and $\widehat{e}_i^{\mathrm{dir}}$ denote the center, covariance, and existence probability produced by direct decoding, while $\widetilde{\mu}_j^{\pi}$, $\widetilde{\Sigma}_j^{\pi}$, and $\widetilde{e}_j^{\pi}$ denote their composed-path counterparts. The three terms therefore measure disagreement in position, covariance geometry, and existence, respectively. Unlike ℓ_{geo} in eq. (37), which compares a prediction with a future target and uses binary cross-entropy for target existence supervision, D_{match} compares two predictions and therefore uses a symmetric squared difference for their existence probabilities.

Let $\mathcal{M}_{\Delta,\pi}$ denote the set of motion groups for which valid direct–composed correspondence is available at horizon Δ, and let $\Pi_{\Delta,\pi}^{G,\mathrm{dc}}$ denote the corresponding Gaussian-level direct–composed transport plan. We define the geometry-level composition loss

$$\mathcal{L}_{\mathrm{gcomp}} = \mathbb{E}_{\Delta,\pi}\left[\sum_{m\in\mathcal{M}_{\Delta,\pi}} d_{\mathrm{SE}(3)}\left(\widehat{T}_{m,t\to t+\Delta}^{\mathrm{dir}}, \widehat{T}_{m,t\to t+\Delta}^{\pi}\right) + \lambda_{\mathrm{gc}} D_{\mathrm{match}}\left(\widehat{\mathcal{G}}_{t+\Delta}^{\mathrm{dir}}, \widetilde{\mathcal{G}}_{t+\Delta}^{\pi}; \Pi_{\Delta,\pi}^{G,\mathrm{dc}}\right)\right]. \tag{43}$$

The first term compares the *motion-group-level transformations*: it penalizes disagreement between the rigid motion decoded directly over the full horizon and the motion obtained by composing shorter decoded transformations. The second term compares the resulting *Gaussian-level scene geometry*: it penalizes disagreement in matched centers, covariances, and existence probabilities. The coefficient $\lambda_{\mathrm{gc}} \geq 0$ controls the relative strength of this Gaussian-level consistency term.

Thus, $\mathcal{L}_{\mathrm{gcomp}}$ in eq. (11) is the geometric counterpart of the latent composition loss $\mathcal{L}_{\mathrm{comp}}$ in eq. (8): $\mathcal{L}_{\mathrm{comp}}$ aligns direct and composed paths in predictive latent space, whereas $\mathcal{L}_{\mathrm{gcomp}}$ encourages their decoded motion representations and Gaussian geometry to be path-consistent. This should be distinguished from the target-anchored geometry loss $\mathcal{L}_{\mathrm{geo}}$ in eq. (39), which pulls each prediction route independently towards the future target geometry.

In the idealized zero-error case, if both the direct and composed decoded states exactly match the same target geometry under unambiguous correspondence, their Gaussian-level endpoint geometries necessarily agree, and the corresponding D_{match} term in eq. (42) becomes redundant. Its role is therefore not to provide an additional notion of target correctness, but to regularize the finite-error regime encountered in practice, where imperfect prediction, soft correspondence, noisy target tracking, and finite model capacity can leave two individually accurate paths with different residual errors[2]. Direct–composed supervision explicitly suppresses this path-dependent discrepancy.

Moreover, agreement in endpoint Gaussian geometry does not necessarily imply agreement in the decoder's structured motion decomposition. Because the same endpoint geometry may be produced by different combinations of group-level $\mathrm{SE}(3)$ motion and Gaussian-specific residual deformation, the transformation distance $d_{\mathrm{SE}(3)}$ in eq. (41), as used in eq. (11), additionally encourages the underlying group transformations themselves to compose consistently across temporal paths. The target-anchored losses $\mathcal{L}_{\mathrm{end}}$, $\mathcal{L}_{\mathrm{path}}$, and $\mathcal{L}_{\mathrm{geo}}$ in eqs. (6), (7) and (39), respectively, therefore establish predictive accuracy, while $\mathcal{L}_{\mathrm{comp}}$ and $\mathcal{L}_{\mathrm{gcomp}}$ in eqs. (8) and (11) act as complementary latent- and geometry-level path-consistency regularizers.

B.7 Structural dynamics priors

The selective decoder is additionally regularized by soft structural priors that encode simple properties expected of many dynamic scenes. Unlike the target-anchored geometry loss $\mathcal{L}_{\mathrm{geo}}$ in eq. (39), these terms do not compare a prediction with future target geometry. Instead, *they constrain how the decoded geometry is allowed to evolve relative to the causal source state*. They are evaluated for both direct decoded intervals $[t, t+s_j]$ and composed segments $[t+s_{j-1}, t+s_j]$, and are then averaged over sampled horizons and paths.

[2]For example, consider a single Gaussian center with target $\mu^+ = 0$ in one dimension. A direct prediction $\widehat{\mu}^{\mathrm{dir}} = +\varepsilon$ and a composed prediction $\widetilde{\mu}^{\pi} = -\varepsilon$ have the same target error, $|\widehat{\mu}^{\mathrm{dir}} - \mu^+| = |\widetilde{\mu}^{\pi} - \mu^+| = \varepsilon$, yet disagree with each other by $|\widehat{\mu}^{\mathrm{dir}} - \widetilde{\mu}^{\pi}| = 2\varepsilon$. Thus, small and even equal target-relative errors do not imply path agreement away from the zero-error limit.

For a decoded interval $[a, b]$, let $\mathcal{M}_{\text{rigid},a}$ denote motion groups identified from the causal prefix as approximately rigid, and let $\mathcal{E}_{m,a}$ contain selected neighboring pairs of Gaussian primitives within rigid group m. A rigid group should preserve the distances between its constituent Gaussians, so we define

$$\ell_{\text{rigid}}(a, b) = \sum_{m \in \mathcal{M}_{\text{rigid},a}} \sum_{(i,j) \in \mathcal{E}_{m,a}} \left(\|\widehat{\mu}_{i,b} - \widehat{\mu}_{j,b}\|_2 - \|\mu_{i,a} - \mu_{j,a}\|_2 \right)^2 . \tag{44}$$

Here, $\mu_{i,a}$ and $\mu_{j,a}$ are the source centers, while $\widehat{\mu}_{i,b}$ and $\widehat{\mu}_{j,b}$ are the decoded centers obtained through the geometric center update by eq. (10). The rigidity loss ℓ_{rigid} in eq. (44) therefore compares each pairwise distance *before and after prediction*; it is small when the group moves without artificial stretching or compression. This prior complements the group-level $\mathrm{SE}(3)$ transformation defined in eq. (29) by encouraging *approximately rigid groups to remain geometrically coherent after the Gaussian-specific residual displacement* in eq. (10) is added.

The decoder also permits Gaussian-specific non-rigid residuals $\widehat{r}_{i,a\to b}$ through the decoder outputs in eq. (28) and the center update in eq. (10). To prevent neighboring primitives from receiving unnecessarily irregular residual motions, let $\mathcal{E}_a$ denote a set of compatible neighboring Gaussian pairs at time a. We define

$$\ell_{\text{smooth}}(a, b) = \sum_{(i,j) \in \mathcal{E}_a} \kappa_{ij} \, \|\widehat{r}_{i,a\to b} - \widehat{r}_{j,a\to b}\|_2^2 , \tag{45}$$

where $\kappa_{ij} \geq 0$ is a compatibility weight derived from geometric proximity, semantic similarity, or estimated motion similarity. Hence, the smoothness loss ℓ_{smooth} in eq. (45) encourages strongly compatible neighboring Gaussians to have similar local residual displacements, while allowing weakly related neighbors to deform more independently. Importantly, eq. (45) regularizes only the *non-rigid residual* $\widehat{r}_{i,a\to b}$ rather than the complete motion, so the coherent group-level rigid transformation in eq. (29) is not penalized.

Finally, let $\mathcal{I}_{\text{static},a}$ denote Gaussian primitives identified from the causal prefix as belonging to approximately static regions. Their predicted centers, obtained from eq. (10), should remain close to their source locations, giving

$$\ell_{\text{static}}(a, b) = \sum_{i \in \mathcal{I}_{\text{static},a}} \|\widehat{\mu}_{i,b} - \mu_{i,a}\|_2^2 . \tag{46}$$

Thus, the static-region loss ℓ_{static} in eq. (46) directly penalizes drift of geometry that the available causal evidence indicates should remain stationary.

Let $\mathcal{L}_{\text{rigid}}$, $\mathcal{L}_{\text{smooth}}$, and $\mathcal{L}_{\text{static}}$ denote the corresponding averages of eqs. (44) to (46), respectively, over the sampled direct intervals and composed segments. Their weighted combination is

$$\mathcal{L}_{\text{struct}} = \lambda_{\text{rigid}} \mathcal{L}_{\text{rigid}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}} + \lambda_{\text{static}} \mathcal{L}_{\text{static}}, \tag{47}$$

where $\lambda_{\text{rigid}}, \lambda_{\text{smooth}}, \lambda_{\text{static}} \geq 0$ control the strengths of rigidity preservation, local residual smoothness, and static-region stability, respectively. In summary, $\mathcal{L}_{\text{rigid}}$ derived from eq. (44) discourages deformation within approximately rigid groups, $\mathcal{L}_{\text{smooth}}$ derived from eq. (45) discourages spatially irregular non-rigid residual motion, and $\mathcal{L}_{\text{static}}$ derived from eq. (46) discourages drift of approximately stationary geometry. These three terms are combined into the structural regularizer $\mathcal{L}_{\text{struct}}$ in eq. (47).

These terms are soft priors rather than universal physical constraints. The sets $\mathcal{M}_{\text{rigid},a}$ and $\mathcal{I}_{\text{static},a}$, as well as the compatibility weights κ_{ij} appearing in eqs. (44) to (46), are determined only from information available in the causal prefix. Rigidity and static penalties in eqs. (44) and (46) are weakened or disabled when prefix-derived evidence indicates non-rigid deformation, independent motion, or topology change. Thus, the structural regularizer in eq. (47) constrains decoded dynamics without introducing future information into the prediction branch.

B.8 Representation regularization and optimization

An EMA target does not by itself preclude all degenerate representation solutions. We therefore apply variance and covariance regularization to scene-, motion-group-, and Gaussian-level student embeddings, following the general principle of VICReg (Bardes et al., 2022). We write

$$\mathcal{L}_{\text{rep}} = \mathcal{L}_{\text{var}} + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}}, \tag{48}$$

where $\mathcal{L}_{\text{var}}$ prevents representation collapse by encouraging each embedding dimension to maintain non-trivial variation across the training batch, while $\mathcal{L}_{\text{cov}}$ discourages different embedding dimensions

from carrying redundant information by penalizing off-diagonal covariance. The coefficient $\lambda_{\text{cov}} \geq 0$ controls the relative strength of the covariance penalty. These regularizers are applied separately at the scene, motion-group, and Gaussian levels of the hierarchical state in eq. (4). Unlike the predictive losses above, $\mathcal{L}_{\text{rep}}$ does not specify what future state should be predicted; rather, it constrains the representation itself so that predictive agreement cannot be achieved through a collapsed or highly redundant latent space. The precise hierarchy-level definitions are given in section F.1.2.

The generic training procedure is summarized in Algorithm 1. The controlled experiments in section 5 instantiate the corresponding ablation objectives and optimize them directly from initialization rather than using a staged training curriculum. Exact model, optimization, regularization, checkpoint-selection, and experiment-specific objective settings are provided in section F.1.

C Hybrid Correspondence Details

Recall from eq. (2) that a Gaussian scene is represented as a set of primitives. Its raw storage order does not define canonical identity or correspondence. For example, the same three persistent primitives may be stored as

$$[A, B, C] \qquad \longrightarrow \qquad [C, A, B], \tag{49}$$

without any change in the represented scene. Naive index matching would nevertheless compare A with C, B with A, and C with B. We therefore distinguish a primitive's *raw array position* from its *persistent canonical identity*:

$$\text{raw array position} \;\neq\; \text{persistent canonical identity}. \tag{50}$$

This distinction directly affects predictive learning. If two equivalent Gaussian states use different primitive orderings, an index-wise latent or geometry loss may penalize an otherwise correct prediction. More generally, a Gaussian set is not uniquely parameterized: densification may split one primitive into several, pruning may remove primitives, and independent reconstructions may reorder, merge, or relocate primitives while representing nearly the same scene.

The correspondence problem becomes still more general under topology-changing evolution. A primitive may split, several primitives may merge, a primitive may disappear, or newly represented geometry may appear:

$$\begin{aligned} A &\longrightarrow \{A_1, A_2\}, && \text{one-to-many split,} \\ \{A, B\} &\longrightarrow C, && \text{many-to-one merge,} \\ A &\longrightarrow \varnothing, && \text{death / disappearance,} \\ \varnothing &\longrightarrow F, && \text{birth / appearance.} \end{aligned}$$

Correspondence is therefore not always one-to-one: it may be one-to-many, many-to-one, one-to-zero, or zero-to-one. A raw index-wise comparison between arbitrary dynamic Gaussian states is consequently not well defined.

C.1 Prefix-anchored canonical bank

Our primary representation uses a prefix-anchored canonical Gaussian bank

$$\mathcal{G}_t^{\text{can}} = \{g_{i,t}^{\text{can}}\}_{i=1}^{N}, \tag{51}$$

constructed using only observations available up to the context endpoint t. Here, $i \in \{1, \ldots, N\}$ denotes a *persistent canonical slot identifier*, and N is the fixed slot capacity of the bank, including both active primitives and reserved inactive slots. This differs from both the raw array position discussed in eq. (50) and N_t in eq. (2), which denotes the number of active scene primitives at time t. Persistent canonical identity therefore means that a slot identifier is deliberately preserved across time while its correspondence remains valid; it does not require the Gaussian geometry associated with that slot to remain fixed. When this canonical-bank representation is used, the active Gaussian scene $\mathcal{G}_t$ in eq. (2) is the active subset of $\mathcal{G}_t^{\text{can}}$, while the remaining slots are reserved inactive slots available for future activation.

For target construction, the canonical bank $\mathcal{G}_t^{\mathrm{can}}$ is cloned at the context endpoint t and tracked forward using only observations available up to each future target time, following the causal protocol in eq. (18). Its parameters at or before t are never retroactively revised using future observations. Within each causal prefix, any densification or refinement used to construct the canonical bank must therefore be completed before target tracking; future observations are not allowed to alter the bank or its states at or before t. Ordinary future evolution is then represented through deformation and existence of persistent slots rather than uncontrolled changes in indexing.

Newly represented geometry may be handled by a reserved pool of inactive canonical slots, consistent with the decoder-side set $\mathcal{I}_a^{\mathrm{dec}}$ introduced in eq. (28). Reserved slots may carry motion-group or spatial anchors that provide an initial prior for future activation. When activated, the predicted local residual $\widehat{r}_{i,a\to b}$, deformation $\widehat{U}_{i,a\to b}$, and existence probability $\widehat{e}_{i,b}$ determine the slot's future geometry through eqs. (10) and (32). The forward target tracker may similarly associate emerging target geometry with available inactive slots using causal group, spatial, or semantic affinity before invoking residual OT. Persistent canonical identities therefore provide a stable causal bookkeeping mechanism without relying on identities obtained by fitting the complete trajectory.

A persistent match is treated as valid only when the two states carry the same persistent canonical identity and that identity remains unique and valid in both states. For two Gaussian states $\mathcal{G}$ and $\mathcal{G}'$, let

$$\mathcal{P}_{\mathrm{pers}}(\mathcal{G},\mathcal{G}') = \left\{ (i,j) : \begin{array}{l} \text{slots } i \text{ and } j \text{ share the same persistent} \\ \text{canonical identity and the match remains unique} \end{array} \right\}. \tag{52}$$

For prediction–target comparison, validity is supplied by the forward target tracker; for direct–composed comparison, it is inherited from the common prefix-anchored canonical bank. A split, merge, canonical reassignment, or otherwise uncertain correspondence invalidates the corresponding hard one-to-one match, and that mass is instead passed to residual matching. Temporary occlusion need not invalidate canonical identity when the causal tracker can maintain the association through the occluded interval.

C.2 Motion-group correspondence

Primitive-level identities may become uncertain under local refinement, occlusion, split, or merge events. Motion groups therefore provide a more stable intermediate unit of correspondence. Recall that $m(i)$ denotes the motion group associated with Gaussian i. In the general formulation, groups may be formed and tracked causally using geometry, motion, semantics, or their combination.

The group-level correspondence matrix Π^O used in eq. (5) records correspondence between motion groups in the two states under comparison. Persistent group identities give hard correspondences when available, while uncertain cases may receive soft correspondence mass. Retaining a motion-group identity even when its supporting Gaussian primitives change allows the group-level representation to remain stable through local split, merge, birth, and death events.

To expose this information to Gaussian-level matching, let

$$\kappa_{mn}^{O} \in [0,1] \tag{53}$$

denote the compatibility between group m in the first state and group n in the second. Persistent group identity gives $\kappa_{mn}^{O} \in \{0,1\}$ as a hard special case, while uncertain group correspondence may produce intermediate values derived from normalized correspondence mass in Π^O. Gaussian primitives belonging to strongly corresponding groups are therefore preferred without requiring group compatibility itself to be binary.

C.3 Hybrid Gaussian correspondence

At the Gaussian level, valid persistent canonical identities are applied first. Let $\Pi^{G,\mathrm{pers}}$ denote a partial transport plan supported only on the valid persistent pairs in eq. (52). For source and target marginal masses $\mathbf{p}$ and $\mathbf{q}$, respectively, the persistent plan satisfies

$$\Pi_{ij}^{G,\mathrm{pers}} = 0 \quad \text{for} \quad (i,j) \notin \mathcal{P}_{\mathrm{pers}}, \qquad \Pi^{G,\mathrm{pers}}\mathbf{1} \le \mathbf{p}, \qquad (\Pi^{G,\mathrm{pers}})^{\top}\mathbf{1} \le \mathbf{q}. \tag{54}$$

Only the remaining unmatched mass is passed to OT. Define the residual marginals

$$\mathbf{p}_{\mathrm{res}} = \mathbf{p} - \Pi^{G,\mathrm{pers}}\mathbf{1}, \qquad \mathbf{q}_{\mathrm{res}} = \mathbf{q} - (\Pi^{G,\mathrm{pers}})^{\top}\mathbf{1}. \tag{55}$$

The residual correspondence plan $\Pi^{G,\mathrm{res}}$ operates only on this remaining mass. The complete Gaussian correspondence is therefore

$$\Pi^{G} = \Pi^{G,\mathrm{pers}} + \Pi^{G,\mathrm{res}}, \tag{56}$$

where persistent and residual correspondence operate on disjoint mass. Equivalently, the matching hierarchy is

$$\text{persistent canonical identity} \longrightarrow \text{motion-group compatibility} \longrightarrow \text{OT for residual ambiguity.} \tag{57}$$

To define the residual OT stage independently of the particular two Gaussian states being compared, consider arbitrary states

$$\mathcal{G} = \{g_i\}_{i=1}^{N}, \qquad \mathcal{G}' = \{g'_j\}_{j=1}^{N'}.$$

Let $m(i)$ and $m'(j)$ denote the motion groups of primitive i in $\mathcal{G}$ and primitive j in $\mathcal{G}'$, respectively. For a residual source primitive i and residual target primitive j, we use the generic cost

$$C_{ij}(\mathcal{G},\mathcal{G}') = \beta_\mu \|\mu_i - \mu'_j\|_2^2 + \beta_f d_{\cos}(f_i, f'_j) + \beta_\Sigma d_{\mathrm{SPD}}(\Sigma_i, \Sigma'_j) + \beta_O \left(1 - \kappa^{O}_{m(i),m'(j)}\right), \tag{58}$$

where μ_i and μ'_j are the Gaussian centers, Σ_i and Σ'_j are their covariances, and f_i and f'_j are semantic features when available. The semantic term is omitted when such features are unavailable. The function $d_{\cos}$ is the cosine distance defined in section 3.2, while d_{SPD} is the SPD covariance distance used in eq. (37). The non-negative coefficients $\beta_\mu, \beta_f, \beta_\Sigma, \beta_O$ control the contributions of position, semantics, covariance geometry, and motion-group compatibility, respectively.

Thus, $C_{ij}(\mathcal{G},\mathcal{G}')$ is small when two primitives are spatially close, semantically compatible, geometrically similar, and associated with strongly corresponding motion groups. Hard group identity is recovered when κ^O is binary, whereas soft group correspondence yields a graded compatibility penalty.

Rather than forcing a single lowest-cost assignment, we obtain a soft residual correspondence through entropically regularized optimal transport following the Lagrangian Sinkhorn formulation[3] of Cuturi (2013, Sec. 4, Eq. (2)).

$$\Pi^{G,\mathrm{res}} = \arg\min_{\Pi \in \mathcal{U}(\mathbf{p}_{\mathrm{res}}, \mathbf{q}_{\mathrm{res}})} \left\{ \langle \Pi, C \rangle + \varepsilon_{\mathrm{OT}} \sum_{i,j} \Pi_{ij} \left(\log \Pi_{ij} - 1 \right) \right\}, \tag{59}$$

where $\mathcal{U}(\mathbf{p}_{\mathrm{res}}, \mathbf{q}_{\mathrm{res}})$ is the set of non-negative transport plans with residual marginals $\mathbf{p}_{\mathrm{res}}$ and $\mathbf{q}_{\mathrm{res}}$, and $\varepsilon_{\mathrm{OT}} > 0$ controls entropy regularization. The first term favors low-cost geometric, semantic, and group-compatible matches, while the entropy term permits soft correspondence when identity is ambiguous. We solve eq. (14) with Sinkhorn iterations after fixing all valid persistent correspondences.

A dustbin row and column augment the residual transport problem in eq. (14) to absorb genuinely unmatched mass. When dustbins are present, $\mathbf{p}_{\mathrm{res}}$ and $\mathbf{q}_{\mathrm{res}}$ in eq. (14) denote the corresponding augmented residual marginals, including their dustbin masses. The target-side dustbin column represents predicted primitives that cannot be matched to represented future geometry, corresponding to disappearance, deactivation, or otherwise unmatched predicted slots; this is the same dustbin $\varnothing$ used to construct the target existence value in eq. (38). Conversely, the predicted-side dustbin row absorbs target geometry that cannot be assigned to an existing or reserved inactive canonical slot. Newly represented geometry is first allowed to match an available inactive slot, so the dustbin acts as a fallback for genuinely unmatched mass rather than the primary birth mechanism.

Because $\Pi^{G,\mathrm{res}}$ in eq. (14) is soft, it permits fractional mass assignments compatible with non-one-to-one correspondence. One predicted primitive may distribute residual mass over several target primitives during a split, while several predicted primitives may place residual mass on the same target lineage during a merge. This flexibility does not by itself identify the true lineage; rather, it avoids imposing an incorrect one-to-one correspondence when the available geometric, semantic, and

[3] Cuturi writes the regularized objective as $\langle P, M \rangle - \lambda^{-1} h(P)$ with $h(P) = -\sum_{ij} P_{ij} \log P_{ij}$. Setting $\varepsilon_{\mathrm{OT}} = \lambda^{-1}$ gives the form used here up to $-\varepsilon_{\mathrm{OT}} \sum_{ij} P_{ij}$, which is constant under fixed transport marginals and therefore does not change the optimizer.

group evidence is ambiguous. Persistent matches remain fixed, and only residual unmatched mass participates in this soft matching.

The experiment-specific instantiation of the residual transport problem, including its count-aware marginals, dustbin cost, Sinkhorn settings, and event definitions, is given in section F.4.1.

C.4 Correspondence for latent and decoded states

The correspondence mechanism is defined from the associated Gaussian slots, canonical identities, motion-group information, and Gaussian geometry; it is not inferred from the latent discrepancy that it is subsequently used to evaluate. Consequently, when the Gaussian-level latent distance in eq. (5) compares tokens z_i^G and $z_j'^G$, the correspondence matrix Π^G is inherited from the associated Gaussian states or canonical slot assignments. This avoids a circular construction in which latent similarity would determine the same correspondence under which latent similarity is measured.

The same hybrid construction is used whenever two Gaussian sets require correspondence. For prediction–target matching,

$$(\mathcal{G}, \mathcal{G}') = \left(\widehat{\mathcal{G}}, \mathcal{G}^+\right), \tag{60}$$

yielding Π^G as used in eqs. (5) and (37). For direct–composed matching,

$$(\mathcal{G}, \mathcal{G}') = \left(\widehat{\mathcal{G}}^{\mathrm{dir}}, \widetilde{\mathcal{G}}^{\pi}\right), \tag{61}$$

yielding the analogous plan $\Pi^{G,\mathrm{dc}}$ used in eq. (42). The correspondence machinery is therefore shared; only the two Gaussian states supplied to it differ.

By default, the resulting correspondence plans Π^G and $\Pi^{G,\mathrm{dc}}$ are treated as matching assignments when used inside the predictive and geometric objectives: gradients from those losses are not propagated through the construction of the correspondence plans. This separates optimization of the predictive model from optimization of the matching rule and prevents the model from reducing a prediction loss merely by altering the correspondence itself. A differentiable Sinkhorn matcher could instead be used when end-to-end optimization of correspondence is desired.

Persistent canonical identity provides useful correspondence when it remains valid, but Gaussian predictions should not be rewarded or penalized for reproducing an arbitrary storage ordering. The resulting comparison is therefore invariant to raw array permutation while remaining sensitive to reliable persistent canonical correspondence. When such identity becomes uncertain or invalid, the hybrid mechanism in eq. (13) progressively falls back to residual OT rather than imposing an incorrect index-wise match.

D Temporal Composition: Analysis

D.1 Intuition and operator interpretation

The temporal-composition objective $\mathcal{L}_{\mathrm{comp}}$ in eq. (8) enforces a simple principle: if two admissible prediction routes describe the same scene evolution over the same elapsed time, they should produce compatible predictive states. For a cumulative horizon s_j, the model may reach time $t + s_j$ either directly through $\widehat{Z}^{\mathrm{dir}}_{t+s_j}$ in eq. (26) or recursively through $\widetilde{Z}^{\pi}_{t+s_j}$ in eq. (25). Temporal composition penalizes disagreement between these routes.

For example, the same endpoint $t + 6$ may be reached as

$$t \longrightarrow t+6, \qquad t \longrightarrow t+2 \longrightarrow t+6, \qquad t \longrightarrow t+4 \longrightarrow t+6.$$

The intermediate states differ, but all routes describe the same total evolution. The objective therefore discourages the horizon-conditioned transition Φ_ϕ in eq. (21) from learning mutually incompatible shortcuts for different temporal decompositions.

At the operator level, the controlled composition law is given in eq. (27). In the action-free, time-homogeneous case it reduces to

$$\Phi_\phi(Z_t, \delta_1 + \delta_2) \approx \Phi_\phi(\Phi_\phi(Z_t, \delta_1), \delta_2)\,.$$

This is an approximate temporal-composition property rather than an assumption of exact semigroup structure.

The two stop-gradient directions in $\mathcal{L}_{\text{comp}}$ implement symmetric two-way agreement: in one term the composed route serves as a fixed target while the direct route receives gradients, and in the other the roles are reversed. The stop-gradient operator therefore changes gradient flow during optimization but not the numerical discrepancy between the two predictions.

Importantly, composition consistency does not establish target correctness. Direct and composed predictions may agree while both are inaccurate. The endpoint loss $\mathcal{L}_{\text{end}}$ in eq. (6) and path loss $\mathcal{L}_{\text{path}}$ in eq. (7) provide future-target anchoring, whereas $\mathcal{L}_{\text{comp}}$ specifically reduces dependence on the temporal route used to reach the future state.

D.2 Proof of Proposition 1

Proof. From $\ell^{\pi}_{\text{path},j} = 0$ and $\omega_j > 0$, the local path-supervision contribution inherited from eq. (7) gives

$$D_{\text{hier}}\Big(\widetilde{Z}^{\pi}_{t+s_j}, Z^{+}_{t+s_j}\Big) = 0.$$

Hence

$$\widetilde{Z}^{\pi}_{t+s_j} = Z^{+}_{t+s_j}$$

on the supervised matched support.

Next, $\ell^{\pi}_{\text{comp},j} = 0$. By its definition as the local contribution of eq. (8), both constituent terms are non-negative. Since $u_j + \nu_j > 0$, at least one direct–composed distance has positive weight and must therefore vanish. Because stop-gradient changes only gradient flow and not numerical values,

$$D_{\text{hier}}\Big(\widehat{Z}^{\text{dir}}_{t+s_j}, \widetilde{Z}^{\pi}_{t+s_j}\Big) = 0.$$

Therefore

$$\widehat{Z}^{\text{dir}}_{t+s_j} = \widetilde{Z}^{\pi}_{t+s_j}$$

on the supervised matched support.

Combining the two equalities proves proposition 1. □

D.3 Finite-error rollout accumulation

Proposition 1 describes the idealized zero-loss case. We now derive the finite-error rollout bound stated in eq. (9).

For this analysis, identify the supervised matched components of the hierarchical predictive state with their concatenated finite-dimensional representation, and let $\|\cdot\|$ denote a fixed norm on this representation. For segment j of a temporal partition $\pi = (\delta_1, \ldots, \delta_J)$, define

$$\Phi_j(Z) = \Phi_\phi\big(Z, \delta_j, A_{t+s_{j-1}:t+s_j-1}\big)\,,$$

with the action argument omitted in passive forecasting.

Let

$$Z^{\star}_0, Z^{\star}_1, \ldots, Z^{\star}_J$$

denote a reference trajectory, with

$$Z^{\star}_j = Z^{+}_{t+s_j}, \qquad j = 1, \ldots, J,$$

at the supervised future horizons. The initial reference state $Z^{\star}_0$ need not equal the realized predictive state Z_t.

Assume that, on the relevant state region, each segment transition Φ_j is locally L_j-Lipschitz,

$$\|\Phi_j(Z) - \Phi_j(Z')\| \le L_j \|Z - Z'\|,$$

and that its local approximation error when applied to the reference trajectory is bounded by

$$\big\|\Phi_j(Z^{\star}_{j-1}) - Z^{\star}_j\big\| \le \epsilon_j.$$

Define the realized recurrent prediction error after segment j as

$$e_j = \left\| \widetilde{Z}^{\pi}_{t+s_j} - Z^{\star}_j \right\|,$$

with initial mismatch

$$e_0 = \|Z_t - Z^{\star}_0\|.$$

Using the recursive rollout in eq. (25),

$$\begin{aligned} e_j &= \left\| \Phi_j\left(\widetilde{Z}^{\pi}_{t+s_{j-1}}\right) - Z^{\star}_j \right\| \\ &\le \left\| \Phi_j\left(\widetilde{Z}^{\pi}_{t+s_{j-1}}\right) - \Phi_j(Z^{\star}_{j-1}) \right\| + \left\| \Phi_j(Z^{\star}_{j-1}) - Z^{\star}_j \right\| \\ &\le L_j e_{j-1} + \epsilon_j. \end{aligned}$$

Applying this recurrence repeatedly gives

$$\begin{aligned} e_J &\le L_J e_{J-1} + \epsilon_J \\ &\le L_J L_{J-1} e_{J-2} + L_J \epsilon_{J-1} + \epsilon_J \\ &\le \cdots \\ &\le \left(\prod_{q=1}^{J} L_q \right) e_0 + \sum_{j=1}^{J} \left(\prod_{q=j+1}^{J} L_q \right) \epsilon_j. \end{aligned}$$

Since $Z^{\star}_J = Z^{+}_{t+s_J}$, this is exactly the bound in eq. (9):

$$\left\| \widetilde{Z}^{\pi}_{t+s_J} - Z^{+}_{t+s_J} \right\| \le \left(\prod_{q=1}^{J} L_q \right) e_0 + \sum_{j=1}^{J} \left(\prod_{q=j+1}^{J} L_q \right) \epsilon_j.$$

Here, an empty product is defined as one. If the reference trajectory is initialized at the realized predictive state, $Z^{\star}_0 = Z_t$, then $e_0 = 0$ and the initial-mismatch term vanishes.

The bound separates three effects. First, e_0 measures any mismatch between the rollout initialization and the chosen reference trajectory. Second, the local errors ϵ_j measure how accurately the learned transition propagates each reference state across one temporal segment. Third, the products of Lipschitz factors quantify how errors introduced earlier in the rollout can be amplified by subsequent applications of Φ_ϕ.

The path loss $\mathcal{L}_{\text{path}}$ in eq. (7) does not directly minimize the local quantities ϵ_j. Instead, it supervises the realized recurrent states encountered when Φ_ϕ is applied to its own predictions and therefore primarily addresses recursive target accuracy. Temporal composition $\mathcal{L}_{\text{comp}}$ in eq. (8) constrains the discrepancy between those recurrent states and their corresponding direct predictions and therefore primarily addresses path dependence. Experiment 1 empirically separates these complementary effects.

E Training Procedure

Algorithm 1 summarizes the generic 4DGS-JEPA training procedure. For clarity, one terminal horizon Δ and one temporal partition π are sampled per minibatch; multiple samples simply give a Monte Carlo estimate of the corresponding expectations in the full objective.

For passive forecasting, the action arguments in Algorithm 1 are omitted. Correspondences used by the latent and geometric losses are constructed from persistent canonical identity and Gaussian-level matching rather than from the latent discrepancy itself. In the current controlled experiments, Experiments 1–2 use fixed persistent correspondence, while Experiment 3 evaluates residual OT independently and contains no learned transition; hence the reported experiments do not backpropagate through Sinkhorn matching.

Experiment 1 and Experiment 2 instantiate restricted versions of Algorithm 1. Experiment 1 applies target-anchored geometry only to the direct route, selectively enables $\mathcal{L}_{\text{path}}$ and $\mathcal{L}_{\text{comp}}$, and disables $\mathcal{L}_{\text{gcomp}}$. Experiment 2 omits $\mathcal{L}_{\text{path}}$, supervises both decoded routes at cumulative endpoints, and varies the group- and Gaussian-level components of $\mathcal{L}_{\text{gcomp}}$. Experiment 3 contains no optimization and therefore does not instantiate Algorithm 1.

Algorithm 1 Temporally compositional 4DGS-JEPA training

Require: Observations $X_{1:T}$; horizon set $\mathcal{K}$; causal builders $B_{\mathrm{on}}, B^{+}_{\mathrm{on}}$
Require: Student encoder E_θ; EMA target encoder $E_{\bar\theta}$; transition Φ_ϕ; decoder $Q^{\mathcal{G}}_\psi$

for each minibatch of context endpoints t **do**
  $\mathcal{G}_{1:t} \leftarrow B_{\mathrm{on}}(X_{1:t})$ ▷ causal Gaussian prefix
  $Z_t \leftarrow E_\theta(\mathcal{H}_t)$
  Sample $\Delta \in \mathcal{K}$ and $\pi = (\delta_1, \dots, \delta_J)$ with $\sum_{j=1}^{J} \delta_j = \Delta$
  $\mathcal{G}^{+}_{t+1:t+\Delta} \leftarrow B^{+}_{\mathrm{on}}(\mathcal{G}_t, X_{t+1:t+\Delta})$ ▷ forward-only targets
  $\widetilde{Z}^{\pi}_t \leftarrow Z_t$; $\widetilde{\mathcal{G}}^{\pi}_t \leftarrow \mathcal{G}_t$
  for $j = 1, \dots, J$ **do**
    $s_j \leftarrow \sum_{q=1}^{j} \delta_q$
    $Z^{+}_{t+s_j} \leftarrow \mathrm{stopgrad}\Big(E_{\bar\theta}(\mathcal{H}^{+}_{t+s_j})\Big)$
    $\widehat{Z}^{\mathrm{dir}}_{t+s_j} \leftarrow \Phi_\phi(Z_t, s_j, A_{t:t+s_j-1})$
    $\widetilde{Z}^{\pi}_{t+s_j} \leftarrow \Phi_\phi(\widetilde{Z}^{\pi}_{t+s_{j-1}}, \delta_j, A_{t+s_{j-1}:t+s_j-1})$
    $\widehat{\mathcal{G}}^{\mathrm{dir}}_{t+s_j} \leftarrow Q^{\mathcal{G}}_\psi(Z_t, \widehat{Z}^{\mathrm{dir}}_{t+s_j}, \mathcal{G}_t, s_j)$
    $\widetilde{\mathcal{G}}^{\pi}_{t+s_j} \leftarrow Q^{\mathcal{G}}_\psi(\widetilde{Z}^{\pi}_{t+s_{j-1}}, \widetilde{Z}^{\pi}_{t+s_j}, \widetilde{\mathcal{G}}^{\pi}_{t+s_{j-1}}, \delta_j)$
    Compute direct–target, composed–target, and direct–composed correspondences ▷ section 4
    Accumulate $\mathcal{L}_{\mathrm{path}}$ and $\mathcal{L}_{\mathrm{comp}}$ ▷ eqs. (7) and (8)
    Accumulate target-anchored $\mathcal{L}_{\mathrm{geo}}$ ▷ eq. (39)
    Accumulate applicable structural priors ▷ eq. (47)
  Compute terminal $\mathcal{L}_{\mathrm{end}}$ ▷ eq. (6)
  Compute terminal $\mathcal{L}_{\mathrm{gcomp}}$ ▷ eq. (11)
  Compute hierarchy-level $\mathcal{L}_{\mathrm{rep}}$ ▷ eq. (48)
  Form $\mathcal{L}$ using eq. (12)
  Update (θ, ϕ, ψ) by gradient descent on $\mathcal{L}$
  $\bar\theta \leftarrow \tau\bar\theta + (1-\tau)\theta$ ▷ EMA target update

F Experimental Details and Additional Results

This section provides implementation details, experiment-specific protocol definitions, and additional results for the four controlled experiments in section 5. Experiments 1 and 2 use the same Gaussian-world model architecture and procedural dynamics generator, with every experimental variant trained independently from initialization. Experiment 3 isolates the Gaussian correspondence mechanism and contains no learned model. Experiment 4 returns to predictive dynamics using the same small-model design, augments the controlled Gaussian trajectories with fixed appearance, and evaluates the predicted Gaussian states through an evaluation-only renderer.

Correspondence instantiation across experiments. Experiments 1, 2, and 4 isolate predictive dynamics and temporal composition from correspondence ambiguity. Persistent canonical identities remain valid, primitive correspondence is fixed, and no permutation, split, merge, birth, or death events are introduced. Gaussian-level correspondence therefore reduces to known persistent matching in these forecasting experiments.

Experiment 3 instead isolates the hybrid correspondence mechanism itself. It contains no learned transition Φ_ϕ or model optimization and deliberately perturbs primitive ordering and topology through permutation, split, merge, birth, death, and mixed events. Ground-truth source–target lineage is retained only for evaluation and is hidden from the correspondence procedures.

The correspondence evaluation is invariant to arbitrary raw storage ordering while remaining sensitive to whether transport mass is assigned to valid canonical lineage. Experiment 3 operationalizes this distinction through LineageMass, birth/death detection, and correspondence-induced geometry discrepancy. It thereby tests whether the hybrid rule in eq. (13) preserves reliable persistent canonical identity when available and degrades gracefully toward residual OT as that information is removed.

Compute environment. Experiments 1–3 were executed in Google Colaboratory (Colab) (Edwards et al., 2024) in runtimes provisioned with an NVIDIA A100 GPU. Experiments 1 and 2 use PyTorch with CUDA for model training and evaluation, while Experiment 3 is implemented entirely

Table 6: Controlled Gaussian-world generator used in Experiments 1 and 2.

Quantity	Setting
Motion groups	3
Gaussians per group	6
Sequence length	24
Group-center range	$[-1, 1]^3$
Within-group center std.	0.16
Initial Gaussian scale	$[0.04, 0.10]$
Angular velocity	$[-0.045, 0.045]$
Angular acceleration	$[-0.002, 0.002]$
Translation velocity	$[-0.035, 0.035]^3$
Translation acceleration	$[-0.0015, 0.0015]^3$
Oscillatory translation amplitude	$[0, 0.08]^3$
Oscillatory translation frequency	$[0.08, 0.18]$
Local residual amplitude	$[0, 0.035]$
Local residual frequency	$[0.10, 0.35]$
Log-scale deformation amplitude	$[0, 0.06]^3$

with NumPy-based correspondence computations. Experiment 4 is also implemented in PyTorch; the reported Experiment 4 run used CPU execution. Because these experiments are independent controlled mechanism tests, we do not use runtime differences across them as an evaluation criterion.

F.1 Common controlled setup for Experiments 1 and 2

Experiments 1 and 2 operate directly on procedurally generated anisotropic 3D Gaussian trajectories rather than fitting Gaussian scenes from images. Consequently, the causal builders B_{on} and B^{+}_{on} in eqs. (17) and (18) reduce to causal access to the generated Gaussian states.

Each scene contains $M = 3$ motion groups with 6 Gaussians per group, giving $N = 18$ active primitives. Their evolution follows the structured group-motion and local-deformation construction described in section 5: group trajectories combine linear and accelerated translation and rotation with oscillatory components, while Gaussian-specific residual displacement and anisotropic scale deformation introduce smooth local non-rigidity. Half of the generated scenes contain one static motion group.

Training, validation, and test scenes are independently generated from disjoint scene seeds, so test trajectories are neither reused training scenes nor temporal continuations of them. For Experiments 1 and 2, motion-group and Gaussian identities remain fixed and all active primitives persist throughout each sequence. Consequently, the correspondence matrices in the general formulation reduce to identity alignment in these two controlled experiments. Permutation, split, merge, birth, and death are introduced only in Experiment 3 so that correspondence errors do not confound the temporal- and geometry-composition experiments.

Table 6 reports the parameter ranges used by the generator. Spatial coordinates use a normalized coordinate system and time advances in unit discrete steps. Translation velocities and accelerations are therefore measured in normalized spatial units per step and per step squared, respectively; angular velocities and accelerations are in radians per step and radians per step squared; sinusoidal frequencies are in radians per step; and log-scale deformation amplitudes are dimensionless. The generated dynamics deliberately contain smooth nonlinear effects beyond static persistence or constant-velocity motion while remaining small enough for controlled analysis.

F.1.1 Causal motion-aware encoder

The context length is $L = 4$. In addition to current Gaussian geometry, the encoder receives causal finite-difference motion information. We use

$$v_{i,t} = \mu_{i,t} - \mu_{i,t-1}, \qquad a_{i,t} = \mu_{i,t} - 2\mu_{i,t-1} + \mu_{i,t-2}.$$

Let $\chi(\Sigma) \in \mathbb{R}^6$ denote the log-Cholesky coordinates of an SPD covariance, and let

$$\bar{\mu}_{m,t} = \frac{1}{|\mathcal{I}_m|} \sum_{i \in \mathcal{I}_m} \mu_{i,t}$$

be the current center of motion group m. The primitive input features are

$$x_{i,t} = \left[\frac{\mu_{i,t}}{s_{\mathrm{p}}}, \frac{\mu_{i,t} - \bar{\mu}_{m(i),t}}{s_{\mathrm{r}}}, \mathrm{NormCov}(\Sigma_{i,t}), 1, \mathrm{onehot}(m(i)), \right.$$
$$\left. \left\{ \frac{\mu_{i,t-k} - \mu_{i,t}}{s_{\mathrm{r}}} \right\}_{k=1}^{L-1}, \frac{v_{i,t}}{s_v}, \frac{a_{i,t}}{s_a}, \frac{\chi(\Sigma_{i,t}) - \chi(\Sigma_{i,t-1})}{s_\Sigma} \right],$$

with fixed scales

$$s_{\mathrm{p}} = 1.5, \qquad s_{\mathrm{r}} = 0.20, \qquad s_v = 0.05, \qquad s_a = 0.01, \qquad s_\Sigma = 0.05.$$

For the current covariance coordinate,

$$\mathrm{NormCov}(\Sigma) = \left[\frac{\chi_{1:3}(\Sigma) - \log(0.07)}{0.5}, \frac{\chi_{4:6}(\Sigma)}{0.05} \right].$$

These constants are fixed from the known generator scale and are not estimated from validation or test data.

Current-geometry and motion features are processed by separate MLP branches and fused before hierarchical aggregation. Gaussian tokens are pooled within their known motion groups, and the resulting group tokens are pooled into a scene token. The encoder and horizon-conditioned transition Φ_ϕ each use one Transformer block with four attention heads.

Controlled geometry decoder. Experiments 1 and 2 use a restricted instantiation of the selective decoder in section 3.4. For each motion group, the decoder predicts a 3-D axis–angle rotation increment and a 3-D translation. For each Gaussian, it predicts a 3-D residual displacement together with three diagonal log-scale deformation parameters. The latter define

$$U_i = \mathrm{diag}(\exp(\widehat{s}_i)),$$

which is used in the covariance update. Because every active primitive persists in Experiments 1 and 2, existence is fixed to one and no existence head or birth/death supervision is used. Appearance and semantic attributes are likewise absent from these two controlled forecasting experiments.

Because motion-group and Gaussian identities are fixed, the experimental hierarchical latent distance is the identity-aligned specialization of eq. (5): scene- and group-level discrepancies use cosine distance, while Gaussian-token discrepancies use a SmoothL1/Huber loss with equal hierarchy weights.

Table 7 summarizes the architecture and optimization settings shared by the learned experiments.

F.1.2 Representation regularization

The controlled experiments instantiate the representation regularizer $\mathcal{L}_{\mathrm{rep}}$ introduced in eq. (48) separately at the scene-, motion-group-, and Gaussian-token levels.

Let $V^\ell \in \mathbb{R}^{B_\ell \times d}$ collect the flattened embeddings at hierarchy level $\ell \in \{S, O, G\}$. The variance component is

$$\mathcal{R}_{\mathrm{var}}(V^\ell) = \frac{1}{d} \sum_{q=1}^{d} \left[\max\left(0, 0.5 - \sqrt{\mathrm{Var}(V^\ell_{:,q}) + 10^{-4}}\right) \right]^2,$$

and, writing C^ℓ for the empirical feature covariance,

$$\mathcal{R}_{\mathrm{cov}}(V^\ell) = \frac{1}{d} \sum_{p \neq q} (C^\ell_{pq})^2.$$

Table 7: Common model and optimization settings for Experiments 1 and 2.

Setting	Value
Context length L	4
Train / validation / test scenes	512/32/48
Train / validation / test samples	6656/416/624
Embedding dimension	48
Decoder horizon embedding	16
Transformer layers	1
Attention heads	4
Transformer dropout	0
Batch / evaluation batch size	48/64
Maximum epochs	15
Optimizer	AdamW
Learning rate	$10^{-3} \to 10^{-5}$, cosine
Weight decay	10^{-5}
Gradient clipping	1.0
EMA momentum τ	0.99
Training horizons	$\{1, 2, 4, 8\}$
Independent runs	3

The implementation uses population-form variance in $\mathcal{R}_{\mathrm{var}}$ and the usual sample-covariance normalization for C^ℓ. The hierarchy-level regularizer is

$$\mathcal{R}_{\mathrm{rep}}(Z) = \frac{1}{3} \sum_{\ell \in \{S,O,G\}} \left[\mathcal{R}_{\mathrm{var}}(V^\ell) + 0.05\, \mathcal{R}_{\mathrm{cov}}(V^\ell) \right].$$

At each training step we apply it to both the current student representation and the direct predicted representation,

$$\mathcal{L}_{\mathrm{rep}} = \frac{1}{2} \left[\mathcal{R}_{\mathrm{rep}}(Z_t) + \mathcal{R}_{\mathrm{rep}}(\widehat{Z}^{\mathrm{dir}}_{t+\Delta}) \right].$$

Its outer objective weight is 0.02 in both Experiments 1 and 2.

F.1.3 Checkpoint selection

Every learned run is trained for at most 15 epochs. Checkpoints are selected exclusively using target-relative validation accuracy averaged over the supervised horizons:

$$S_{\mathrm{val}} = \frac{1}{2} \left(\mathrm{FDE}_{\mathrm{dir,val}} + \mathrm{FDE}_{\mathrm{roll,val}} \right).$$

The direct and rollout terms are averaged over the validation samples and supervised horizons $\{1, 2, 4, 8\}$. Validation rollout FDE uses the fixed primary routes

$$(1), \qquad (1, 1), \qquad (2, 2), \qquad (4, 4)$$

for horizons $\Delta = 1, 2, 4, 8$, respectively. No latent- or geometry-composition metric enters checkpoint selection. Models trained with composition losses therefore receive no model-selection advantage on the consistency quantities they explicitly optimize.

F.2 Experiment 1: Temporal Composition

Metric definitions. For target-relative geometry accuracy, we report the *final displacement error*

$$\mathrm{FDE}(\widehat{\mathcal{G}}, \mathcal{G}^+) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\widehat{\mu}_i - \mu_i^+\|_2, \tag{62}$$

using the known persistent correspondences in this experiment. We distinguish *direct FDE*, obtained from $\widehat{\mathcal{G}}^{\mathrm{dir}}_{t+\Delta}$, from *rollout FDE*, obtained from $\widetilde{\mathcal{G}}^{\pi}_{t+\Delta}$. The additional error associated with recursive prediction is summarized by

$$\Delta_{\mathrm{roll}} = \mathrm{FDE}_{\mathrm{roll}} - \mathrm{FDE}_{\mathrm{dir}}. \tag{63}$$

For latent path consistency, we report

$$\mathrm{CompErr}(\Delta,\pi) = D_{\mathrm{hier}}\left(\widehat{Z}^{\mathrm{dir}}_{t+\Delta}, \widetilde{Z}^{\pi}_{t+\Delta}\right), \tag{64}$$

together with the center-level direct–composed geometric discrepancy

$$\mathrm{GeoCompErr}(\Delta,\pi) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \left\|\widehat{\mu}^{\mathrm{dir}}_{i,t+\Delta} - \widetilde{\mu}^{\pi}_{i,t+\Delta}\right\|_2 . \tag{65}$$

Composition metrics are computed only for non-trivial paths with more than one segment; the one-segment case is excluded because the direct and composed computations coincide. We report composition metrics jointly with target-relative FDE because small path discrepancy alone does not imply correct future prediction.

We evaluate supervised horizons $\{1,2,4,8\}$ and unseen interpolation horizons $\{3,6\}$. We additionally distinguish held-out decompositions composed entirely of segment lengths observed during training from a harder setting containing unseen segment lengths. For example, at $\Delta = 8$, training uses $(4,4)$ and $(2,2,4)$, while held-out seen-segment decompositions include $(4,2,2)$, $(2,4,2)$, and $(2,2,2,2)$; unseen-segment tests include $(3,5)$ and $(5,3)$.

Experiment 1 compares the four variants in table 1. All variants share the same data, architecture, EMA teacher, selective decoder, optimization budget, and representation regularization.

As described in the main text, geometry supervision in Experiment 1 is applied only to the direct decoded route. Let $\mathcal{L}^{\mathrm{dir}}_{\mathrm{geo}}$ denote this direct-route restriction of the target-anchored geometry loss in eq. (39). The implemented mean-reduced objective is

$$\mathcal{L}_{\mathrm{Exp1}} = \mathcal{L}_{\mathrm{end}} + 5\mathcal{L}^{\mathrm{dir}}_{\mathrm{geo}} + \mathbf{1}_{\mathrm{path}}\mathcal{L}_{\mathrm{path}} + 0.5\,\mathbf{1}_{\mathrm{comp}}\mathcal{L}_{\mathrm{comp}} + 0.02\mathcal{L}_{\mathrm{rep}},$$

where $\mathbf{1}_{\mathrm{path}}, \mathbf{1}_{\mathrm{comp}} \in \{0,1\}$ specify the corresponding ablation. Thus, the active path-loss coefficient is 1, while the active temporal-composition coefficient is 0.5. Geometry-level composition $\mathcal{L}_{\mathrm{gcomp}}$ in eq. (11) is disabled throughout Experiment 1.

The center component of $\mathcal{L}^{\mathrm{dir}}_{\mathrm{geo}}$ applies the Huber penalty to the Euclidean 3-D center displacement with threshold 0.1. Covariance supervision applies SmoothL1 loss in log-Cholesky coordinates with relative weight 0.25.

At each optimizer step, one terminal horizon is used, cycling uniformly through $\{1,2,4,8\}$ so that all variants receive the same horizon schedule. When multiple training partitions are available for that horizon, one is sampled from the corresponding partition set. The terminal horizons $\{3,6\}$ are never used for training and are evaluated only as unseen-horizon interpolation tests.

Temporal partitions. The training decompositions are

$$\begin{aligned}
\mathcal{D}^{(1)}_{\mathrm{train}} &= \{(1)\},\\
\mathcal{D}^{(2)}_{\mathrm{train}} &= \{(1,1)\},\\
\mathcal{D}^{(4)}_{\mathrm{train}} &= \{(2,2),(1,1,2)\},\\
\mathcal{D}^{(8)}_{\mathrm{train}} &= \{(4,4),(2,2,4)\}.
\end{aligned}$$

The primary evaluation route for each terminal horizon is

$$\begin{aligned}
&\pi^{(1)}_{\mathrm{pri}} = (1), \quad \pi^{(2)}_{\mathrm{pri}} = (1,1), \quad \pi^{(3)}_{\mathrm{pri}} = (1,2),\\
&\pi^{(4)}_{\mathrm{pri}} = (2,2), \quad \pi^{(6)}_{\mathrm{pri}} = (2,4), \quad \pi^{(8)}_{\mathrm{pri}} = (4,4).
\end{aligned}$$

Thus, terminal horizons 3 and 6 are not supervised during training, although the individual segment lengths used in their primary rollouts are observed during training.

To test generalization to unseen decompositions while keeping all segment lengths familiar, we additionally use

$$\begin{aligned}
\mathcal{D}^{(4)}_{\mathrm{held}} &= \{(2,1,1),(1,2,1)\},\\
\mathcal{D}^{(8)}_{\mathrm{held}} &= \{(4,2,2),(2,4,2),(2,2,2,2)\}.
\end{aligned}$$

Table 8: Generalization to temporal decompositions not used during Experiment 1 training, averaged over 3 random seeds. Lower is better.

Method	Held-out, seen segments		Unseen segments	
	CompErr ↓	GeoCompErr ↓	CompErr ↓	GeoCompErr ↓
Endpoint	0.01134 ± 0.00244	0.01428 ± 0.00160	0.00235 ± 0.00047	0.00901 ± 0.00120
Endpoint + Path	0.00286 ± 0.00013	0.01105 ± 0.00146	0.00064 ± 0.00006	$\mathbf{0.00778 \pm 0.00119}$
Endpoint + Composition	0.00454 ± 0.00049	0.01204 ± 0.00125	0.00096 ± 0.00011	0.00803 ± 0.00110
Endpoint + Path + Composition	$\mathbf{0.00260 \pm 0.00009}$	$\mathbf{0.01096 \pm 0.00157}$	$\mathbf{0.00057 \pm 0.00002}$	0.00782 ± 0.00141

A harder evaluation introduces segment lengths 3 and 5, neither of which is used as a transition horizon during training:

$$\mathcal{D}^{(4)}_{\text{unseen}} = \{(1,3),(3,1)\},$$
$$\mathcal{D}^{(8)}_{\text{unseen}} = \{(3,5),(5,3)\}.$$

As shown in table 8, the full path-and-composition model gives the lowest latent path discrepancy in both held-out regimes. CompErr decreases from 0.01134 to 0.00260 for new decompositions made entirely from familiar segment lengths and from 0.00235 to 0.00057 when unseen segment lengths are introduced. This complements the target-relative and composition results reported in section 5.1.

F.3 Experiment 2: Geometry-Level Composition

Metric definitions. Because the controlled generator provides the true motion-group transformations, we evaluate both *structured-motion correctness* and *path agreement*. For this controlled experiment, transformation errors use the rotation–translation surrogate $d^{\text{ctrl}}_{\text{SE(3)}}$ defined in section F.3, rather than the general Lie-group distance in eq. (41). The same controlled distance is used for training the group-level composition term and for the reported transformation metrics. Let $T^{+}_{m,t\to t+\Delta}$ denote the ground-truth relative transformation of motion group m. We report

$$E^{\text{target}}_{T} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} d^{\text{ctrl}}_{\text{SE(3)}}\left(\widehat{T}^{\text{dir}}_{m,t\to t+\Delta}, T^{+}_{m,t\to t+\Delta}\right), \tag{66}$$

and

$$E^{\text{dc}}_{T} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} d^{\text{ctrl}}_{\text{SE(3)}}\left(\widehat{T}^{\text{dir}}_{m,t\to t+\Delta}, \widehat{T}^{\pi}_{m,t\to t+\Delta}\right). \tag{67}$$

At the Gaussian level, we additionally report direct target-relative center FDE from eq. (62), direct covariance error using d_{SPD}, and direct–composed geometry discrepancy D_{match} from eq. (42). These metrics distinguish target accuracy from path consistency.

Experiment 2 uses the same procedural generator and model architecture as Experiment 1, but each variant is trained independently from initialization. All three variants use the same Endpoint + Composition latent objective, data, optimizer, horizon schedule, and temporal-partition schedule. Path supervision $\mathcal{L}_{\text{path}}$ is omitted so that the additional effect of geometry-level composition can be isolated.

For each sampled temporal partition, both the direct and composed decoded states are independently supervised against the same target geometry at every cumulative endpoint. In the implementation, the two route losses are averaged at each endpoint and then averaged over the path. Denoting this mean-reduced controlled version of eq. (39) by $\mathcal{L}^{\text{ctrl}}_{\text{geo}}$, the common objective is

$$\mathcal{L}^{\text{Exp2}}_{\text{base}} = \mathcal{L}_{\text{end}} + 0.5\mathcal{L}_{\text{comp}} + 5\mathcal{L}^{\text{ctrl}}_{\text{geo}} + 0.02\mathcal{L}_{\text{rep}}.$$

Let $\mathcal{L}^{\text{SE(3)}}_{\text{gcomp}}$ and $\mathcal{L}^{\text{match}}_{\text{gcomp}}$ denote the controlled motion-group and Gaussian-level direct–composed terms. The three variants use

$$\mathcal{L}_{\text{latent}} = \mathcal{L}^{\text{Exp2}}_{\text{base}},$$
$$\mathcal{L}_{\text{SE(3)}} = \mathcal{L}^{\text{Exp2}}_{\text{base}} + 0.4\mathcal{L}^{\text{SE(3)}}_{\text{gcomp}},$$
$$\mathcal{L}_{\text{full}} = \mathcal{L}^{\text{Exp2}}_{\text{base}} + 0.4\mathcal{L}^{\text{SE(3)}}_{\text{gcomp}} + 0.5\mathcal{L}^{\text{match}}_{\text{gcomp}}.$$

Table 9: Horizon-wise direct–composed geometry consistency in Experiment 2, averaged over 3 random seeds. Lower is better.

	$E_T^{\mathrm{dc}} \downarrow$			$D_{\mathrm{match}} \downarrow$		
Δ	Latent	+SE(3)	Full	Latent	+SE(3)	Full
2	0.001685±0.000121	**0.000466±0.000089**	0.000528±0.000163	0.000429±0.000056	0.000360±0.000051	**0.000350±0.000056**
4	0.004297±0.000723	**0.001018±0.000259**	0.001102±0.000355	0.001048±0.000167	0.000664±0.000153	**0.000626±0.000082**
8	0.007238±0.000573	**0.002015±0.000311**	0.002345±0.000619	0.001923±0.000152	0.001378±0.000362	**0.001298±0.000124**

The geometry-composition terms are evaluated only at the terminal endpoint of the sampled partition.

The general Method defines a Lie-group transformation discrepancy in eq. (41). Experiment 2 instead uses the controlled rotation–translation surrogate

$$d^{\mathrm{ctrl}}_{\mathrm{SE}(3)}(T_1, T_2) = \sqrt{\theta(R_1^\top R_2)^2 + \|d_2 - d_1\|_2^2 + 10^{-12}}, \qquad T_k = (R_k, d_k),$$

where $\theta(R)$ is the geodesic rotation angle on $\mathrm{SO}(3)$. The implemented group-level composition term is the mean of $d^{\mathrm{ctrl}}_{\mathrm{SE}(3)}$ over motion groups and samples. The same distance is used for the reported target-relative and direct–composed transformation errors E_T^{target} and E_T^{dc} in eqs. (66) and (67).

For Gaussian-level geometry supervision and composition during training, center error uses the Euclidean-displacement Huber penalty with threshold 0.1, while covariance discrepancy uses SmoothL1 distance in log-Cholesky coordinates with relative weight 0.25. Evaluation instead uses the affine-invariant SPD covariance distance. In particular, the direct target-relative covariance error reported in table 2 is

$$E_\Sigma^{\mathrm{target}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} d_{\mathrm{SPD}}\left(\widehat{\Sigma}_i^{\mathrm{dir}}, \Sigma_i^+\right).$$

The reported direct–composed D_{match} uses the same Huber center penalty together with the affine-invariant SPD covariance distance, again with covariance weight 0.25. Because persistent Gaussian identities are fixed, these comparisons are slot-aligned and require no OT. Existence is also fixed to one, so the existence component of the general D_{match} in eq. (42) is identically zero in this experiment.

For numerical stability during recursive Experiment 2 rollouts, decoded covariance matrices are explicitly symmetrized and regularized by adding $10^{-6}I$ before reuse as the source covariance for the next segment.

Because the synthetic generator exposes the absolute motion-group poses at each time, the ground-truth relative transformation can be constructed exactly. This permits both target-relative structured-motion evaluation through E_T^{target} and direct–composed evaluation through E_T^{dc}.

The non-trivial direct–composed geometry evaluation uses horizons $\Delta \in \{2, 4, 8\}$ with primary composed paths $(1, 1)$, $(2, 2)$, and $(4, 4)$, respectively. The one-segment horizon $\Delta = 1$ is excluded from direct–composed metrics because its direct and composed computations coincide.

The ordering in table 9 is stable across $\Delta \in \{2, 4, 8\}$. Relative to latent composition alone, explicit $\mathrm{SE}(3)$ composition reduces E_T^{dc} by approximately 72.4%, 76.3%, and 72.2%, respectively. The full objective reduces D_{match} by approximately 18.5%, 40.2%, and 32.5% at the same horizons. Thus, the motion-group term gives the strongest transform-level consistency, whereas the full objective gives the strongest primitive-level Gaussian consistency.

As a non-degeneracy check at $\Delta = 8$, static persistence has FDE 0.32158, causal constant velocity has FDE 0.07120, and the identity group transformation has $E_T^{\mathrm{target}} = 0.34170$. The three learned variants obtain FDE between 0.06614 and 0.06728 and target-relative structured-transform error between 0.25554 and 0.26124. Their low direct–composed discrepancies therefore cannot be explained by predicting negligible motion.

F.4 Experiment 3: Correspondence Stress Test

Metric definitions. Because the true source–target lineage is known, let $Y_{ij} \in \{0, 1\}$ indicate whether source primitive i and target primitive j share a valid ground-truth lineage relation. For a

Table 10: Experiment 3 configuration.

Setting	Value
Active / reserved primitives	$24/6$
Motion-group labels	4
Semantic dimension	6
Perturbation severities	$\{0.10, 0.25, 0.40\}$
Position-noise std.	0.025
Log-scale-noise std.	0.020
Semantic-noise std.	0.050
Sinkhorn entropy $\varepsilon_{\mathrm{OT}}$	0.08
Maximum Sinkhorn iterations	200
Sinkhorn tolerance	10^{-7}
Real–dustbin cost	1.2
Dustbin–dustbin cost	0
Independent Monte Carlo repetitions	3
Trials per condition and repetition	50
ID-dropout rates	$\{0, 0.10, 0.25, 0.50\}$
Bootstrap samples	2000

predicted correspondence plan Π^G, we report

$$\mathrm{LineageMass} = \frac{\sum_{i,j} Y_{ij} \Pi^G_{ij}}{\sum_{i,j} \Pi^G_{ij} + \varepsilon_{\mathrm{match}}}, \tag{68}$$

where the sums include only real, non-dustbin assignments. The numerator measures correspondence mass assigned to ground-truth-valid lineage edges and the denominator measures total real-to-real correspondence mass. Higher LineageMass therefore indicates greater concentration of predicted matching mass on valid ground-truth lineage relations.

We additionally report birth/death F1 for topology-changing events and the geometry discrepancy induced by the correspondence. The latter measures whether the transport plan pairs geometrically compatible primitives. The count-aware OT construction, event definitions, bootstrap confidence intervals, and additional severity and identity-availability tests are given in section F.4.

Experiment 3 evaluates correspondence independently of the learned transition Φ_ϕ and contains no learned parameters or optimization. Each source scene contains 24 active Gaussian primitives and 6 reserved inactive canonical slots with four motion-group labels. Each primitive additionally carries a 6-dimensional normalized semantic descriptor. We use three independent Monte Carlo repetitions with 50 trials per perturbation type and severity.

For each trial, small target-side center and log-scale perturbations are first applied to the active source primitives. A topology operation then applies permutation, split, merge, death, reserved-slot activation, or a mixture of these events. Split descendants retain the ground-truth lineage of their parent but lose a valid one-to-one persistent ID; merged primitives retain the union of their parent lineages and likewise receive no one-to-one persistent ID. Death removes primitives and birth activates reserved inactive canonical slots. After the topology operation, independent semantic noise is added and each semantic descriptor is renormalized. Permutation and mixed conditions finally randomize the target array ordering.

At perturbation severity r, the nominal number of affected active primitives is

$$n_r = \max(1, \mathrm{round}(r N_{\mathrm{active}})) .$$

F.4.1 Count-aware residual matching

The Method defines the generic residual OT cost C_{ij} in eq. (58). In the controlled correspondence stress test, the covariance component is instantiated through log-scale coordinates, giving

$$C_{ij} = 2\|\mu_i - \mu_j^+\|_2^2 + 0.25\|s_i - s_j^+\|_2^2 + 0.8 d_{\cos}(f_i, f_j^+) + 0.8\, \mathbf{1}[m(i) \neq m(j)],$$

where s_i and s_j^+ are Gaussian log-scale vectors. Because motion-group labels are known exactly in this controlled stress test, the final term is the hard-label specialization of the generic group-compatibility term in eq. (58), with

$$\kappa_{mn}^{O} = \mathbf{1}[m = n].$$

The semantic descriptors are independently perturbed and renormalized, so semantic features cannot serve as perfectly preserved identity labels.

OT only and Hybrid use the same unit-mass convention. After any hard persistent matches have been removed, let n_s and n_t denote the numbers of residual source and target primitives. The dustbin-augmented residual marginals are

$$\mathbf{p}_{\text{res}}^{\text{aug}} = \left(\underbrace{1, \ldots, 1}_{n_s}, n_t \right), \qquad \mathbf{q}_{\text{res}}^{\text{aug}} = \left(\underbrace{1, \ldots, 1}_{n_t}, n_s \right),$$

so that

$$\|\mathbf{p}_{\text{res}}^{\text{aug}}\|_1 = \|\mathbf{q}_{\text{res}}^{\text{aug}}\|_1 = n_s + n_t.$$

The final source row and target column are dustbins. Their capacities equal the cardinality of the opposite residual set, allowing arbitrary source–target cardinality mismatch without imposing a fixed unmatched-mass fraction.

For OT only, no persistent correspondence is removed first, so the complete source and target sets enter this augmented OT problem. For Hybrid, every valid unique persistent canonical correspondence is first assigned one unit of hard mass,

$$\Pi_{ij}^{G,\text{pers}} = 1.$$

Only the remaining unmatched source and target primitives enter the same count-aware OT problem. Persistent and residual correspondence therefore operate on identical mass scales. Split and merge primitives, whose one-to-one persistent identities are invalidated by construction, are handled entirely through residual OT.

The augmented cost matrix assigns cost 1.2 to every real-source–target-dustbin and source-dustbin–real-target entry, while the dustbin–dustbin corner has cost zero. The dustbin cost is fixed and is not learned. If one residual side is empty, each remaining real primitive on the opposite side is assigned one unit of mass directly to or from the corresponding dustbin.

The residual plan $\Pi^{G,\text{res}}$ is obtained from the Sinkhorn problem in eq. (14), using entropy parameter $\varepsilon_{\text{OT}} = 0.08$, at most 200 Sinkhorn iterations, and marginal tolerance 10^{-7}.

F.4.2 Metrics and uncertainty

The primary lineage metric is LineageMass in eq. (68). In the controlled implementation,

$$\varepsilon_{\text{match}} = 10^{-12}.$$

We additionally evaluate the geometry discrepancy induced by the real–real component of the correspondence plan:

$$D_{\text{corr}} = \frac{\sum_{i,j} \Pi_{ij}^{G} \|\mu_i - \mu_j^+\|_2}{\sum_{i,j} \Pi_{ij}^{G} + \varepsilon_{\text{match}}},$$

where dustbin assignments are excluded. Lower D_{corr} indicates that correspondence mass is concentrated on geometrically compatible source–target pairs.

Birth and death are scored separately using precision, recall, and F1. A source primitive is classified as dead when its target-dustbin mass is at least its total mass assigned to real targets. A target primitive is classified as newly born when the mass arriving from reserved source slots together with the source dustbin is at least the mass arriving from active source primitives. Event F1 is the NaN-aware mean of the applicable birth and death F1 scores; an event type absent from the ground truth is treated as not applicable rather than assigned zero.

The main results in table 3 report mean $\pm$ standard deviation across the three Monte Carlo repetition-level aggregates. We additionally compute 95% bootstrap confidence intervals using 2000 resamples

Table 11: Mixed permutation, split, merge, birth, and death perturbations as severity increases, averaged over 3 independent Monte Carlo repetitions.

Severity	Method	LineageMass ↑	Geometry discrepancy ↓
0.10	OT only	0.95694±0.00752	0.05693±0.00157
	Hybrid	**0.96735±0.00620**	**0.05541±0.00108**
0.25	OT only	0.95148±0.00091	0.06502±0.00171
	Hybrid	**0.95988±0.00082**	**0.06349±0.00147**
0.40	OT only	0.91962±0.00344	0.07718±0.00183
	Hybrid	**0.92605±0.00137**	**0.07603±0.00158**

Table 12: Persistent-ID dropout under mixed perturbations at severity 0.25.

Hidden IDs	OT lineage	Hybrid lineage	OT geometry	Hybrid geometry
0	0.95148	**0.95988**	0.06502	**0.06349**
0.10	0.95148	**0.95916**	0.06502	**0.06361**
0.25	0.95148	**0.95838**	0.06502	**0.06376**
0.50	0.95148	**0.95649**	0.06502	**0.06409**

over repetition–trial clusters. A cluster corresponds to one Monte Carlo repetition and trial index, preserving the common replicate structure while aggregating the relevant perturbation conditions. The 95% bootstrap intervals for split/merge LineageMass are $[0.97916, 0.98281]$ for OT only and $[0.98540, 0.98889]$ for Hybrid. The corresponding intervals for aggregate geometry discrepancy are $[0.04977, 0.05043]$ and $[0.04871, 0.04940]$, respectively. These separated intervals are consistent with the small systematic Hybrid advantage in this controlled benchmark; they are not presented as a separate formal hypothesis test.

F.4.3 Mixed perturbations

Table 11 shows the same qualitative behavior at every evaluated severity: Hybrid retains more correspondence mass on valid lineage and produces lower induced geometry discrepancy than OT only. The margin decreases as severity increases because progressively more persistent canonical identities become invalid and a larger fraction of the correspondence problem is delegated to residual OT.

Birth/death identification itself is not the source of the Hybrid advantage. For isolated birth and death perturbations, both OT only and Hybrid attain F1 $= 1$. Under mixed perturbations, averaged across severities, OT only obtains macro event F1 0.8984 and Hybrid obtains 0.8884. The Hybrid advantage is therefore concentrated in lineage preservation and induced geometry correspondence rather than in uniformly improving every auxiliary event metric.

F.4.4 Persistent-ID dropout

To test whether Hybrid depends critically on canonical identity, we hide a fraction of otherwise valid persistent target IDs under mixed perturbations at severity 0.25. Gaussian geometry and ground-truth lineage remain unchanged; only the identity information available to Hybrid is removed. OT only is therefore unchanged across dropout rates.

As shown in table 12, progressively hiding persistent canonical identity causes Hybrid to move smoothly toward the identity-agnostic OT-only solution rather than failing abruptly. This is the intended fallback behavior of the hybrid correspondence principle in eq. (13).

F.5 Experiment 4: Held-Out Future-Frame Rendering

Experiment 4 evaluates whether the Gaussian geometry predicted by 4DGS-JEPA can be converted into a held-out future observation through Gaussian rendering, without adding an image-reconstruction objective to training. The experiment remains fully controlled: Gaussian trajectories are generated directly, RGB color and opacity are fixed over time, and the renderer is applied only after a Gaussian state has been predicted.

F.5.1 Data and held-out-frame protocol

Each trajectory contains 24 frames, 3 motion groups, and 6 Gaussians per group, giving 18 active primitives. The geometric dynamics follow the same structured family used in Experiments 1 and 2: group-level translation and rotation are combined with Gaussian-specific local displacement and anisotropic covariance deformation. Depth-direction motion is reduced relative to the other spatial directions so that the primitives remain stably visible under the fixed camera.

Unlike Experiments 1 and 2, each primitive additionally receives a time-invariant RGB color $c_i \in [0,1]^3$ and opacity $\alpha_i \in [0,1]$. Colors are sampled around group-specific base colors with small primitive-level perturbations, while

$$\alpha_i \sim \mathrm{Unif}(0.55, 0.82).$$

These appearance quantities remain fixed for the complete trajectory and are transported unchanged when a future Gaussian state is decoded. Experiment 4 therefore tests future *geometry-to-image rendering*, not future appearance prediction.

We generate 128 training trajectories, 16 validation trajectories, and 20 test trajectories from mutually disjoint scene seeds. Each trajectory has length $T = 24$. Its final frame is excluded from every training and validation target. With zero-based implementation indexing this is frame 23; in the paper notation it is the terminal frame I_T.

The context length is $L = 4$. Training uses

$$\mathcal{K}_{\mathrm{train}}^{\mathrm{Exp4}} = \{1, 2, 4\},$$

with temporal partitions

$$\begin{aligned}
\mathcal{D}_{\mathrm{train}}^{(1)} &= \{(1)\}, \\
\mathcal{D}_{\mathrm{train}}^{(2)} &= \{(1,1)\}, \\
\mathcal{D}_{\mathrm{train}}^{(4)} &= \{(2,2), (1,1,2)\}.
\end{aligned}$$

The final frame is never contained in any training or validation target under any of these horizons. This gives 2048 training context endpoints and 256 validation context endpoints.

The principal test starts from the last causal context ending at $T-1$ and predicts one step forward,

$$\widehat{Z}_T = \Phi_\phi(Z_{T-1}, 1), \qquad \widehat{\mathcal{G}}_T = Q_\psi^{\mathcal{G}}(Z_{T-1}, \widehat{Z}_T, \mathcal{G}_{T-1}, 1).$$

The held-out target $\mathcal{G}_T$ is used only for evaluation.

A separate non-trivial composition diagnostic starts from time $T-4$ and compares

$$\widehat{Z}_T^{\mathrm{dir}} = \Phi_\phi(Z_{T-4}, 4)$$

with the recurrent route

$$\widetilde{Z}_T^{(2,2)} = \Phi_\phi(\Phi_\phi(Z_{T-4}, 2), 2)\,.$$

The corresponding Gaussian states are decoded directly and recursively as in eqs. (35) and (36).

F.5.2 Model objective and checkpoint selection

Experiment 4 uses the same small hierarchical encoder, horizon-conditioned transition, selective geometry decoder, EMA target encoder, causal motion features, and representation regularization as Experiment 1. The four variants are again Endpoint, Endpoint + Path, Endpoint + Composition, and Endpoint + Path + Composition.

As in Experiment 1, target-anchored geometry supervision is applied only to the direct decoded route. The implemented objective is

$$\mathcal{L}_{\mathrm{Exp4}} = \mathcal{L}_{\mathrm{end}} + 5\mathcal{L}_{\mathrm{geo}}^{\mathrm{dir}} + \mathbf{1}_{\mathrm{path}}\mathcal{L}_{\mathrm{path}} + 0.5\,\mathbf{1}_{\mathrm{comp}}\mathcal{L}_{\mathrm{comp}} + 0.02\mathcal{L}_{\mathrm{rep}}.$$

No RGB, rendering, PSNR, SSIM, or perceptual-image term enters this objective.

Table 13: Experiment 4 controlled data, training, and rendering settings.

Setting	Value
Motion groups	3
Gaussians per group	6
Sequence length	24
Context length	4
Training / validation / test trajectories	128/16/20
Training context endpoints	2048
Validation context endpoints	256
Training horizons	$\{1, 2, 4\}$
Training partitions at $\Delta = 4$	$(2, 2), (1, 1, 2)$
Token dimension	48
Horizon-embedding dimension	16
Transformer blocks / heads	1/4
Batch size	48
Maximum epochs	15
Initial / minimum learning rate	$10^{-3}/10^{-5}$
AdamW weight decay	10^{-5}
EMA momentum	0.99
Gradient clipping	1
Training seeds	$\{0, 1, 2\}$
Image resolution	96×96
Focal length	95 pixels
Camera distance	4.0
Minimum projected variance	0.35
Maximum per-Gaussian alpha	0.995
Background intensity	0.03
RGB / rendering loss	none

At each optimizer step, the terminal horizon cycles uniformly through $\{1, 2, 4\}$. When more than one temporal partition is available, one partition is sampled from the corresponding training set above. Models are independently initialized for each of the three seeds.

Checkpoint selection uses only target-relative Gaussian geometry on the validation trajectories. For each supervised horizon, direct and primary-rollout FDE are evaluated, using primary routes

$$(1), \qquad (1, 1), \qquad (2, 2)$$

for $\Delta = 1, 2, 4$, respectively. The validation score is

$$S_{\text{val}}^{\text{Exp4}} = \frac{1}{2} \left(\text{FDE}_{\text{dir,val}} + \text{FDE}_{\text{roll,val}} \right).$$

No rendering metric or composition discrepancy enters checkpoint selection.

F.5.3 Evaluation-only renderer

For a predicted Gaussian state

$$\widehat{\mathcal{G}} = \{(\widehat{\mu}_i, \widehat{\Sigma}_i, \alpha_i, c_i)\}_{i=1}^{N},$$

we render the image using a fixed perspective camera. The camera is placed at distance 4.0 from the origin and uses focal length 95 pixels for a 96×96 image.

Following standard 3DGS rasterization (Kerbl et al., 2023), each 3D center is projected to the image plane and its covariance is propagated through the local perspective Jacobian. In Experiment 4, a minimum image-plane covariance of $0.35I$ is added for numerical stability. The resulting projected Gaussian contributes

$$a_i(p) = \alpha_i \exp\left[-\frac{1}{2} (p - \mu_i^{2D})^\top (\Sigma_i^{2D})^{-1} (p - \mu_i^{2D}) \right],$$

with contributions truncated outside the region whose squared Mahalanobis distance exceeds 18. Alpha values are clipped at 0.995, and Gaussians are composited front-to-back. The background RGB intensity is fixed to 0.03.

Table 14: Experiment 4 held-out next-frame results for the learned variants, averaged over 20 test trajectories and summarized across 3 independently trained seeds.

Method	FDE ↓	Cov. err. ↓	MSE ↓	PSNR ↑	SSIM ↑
Endpoint	0.009720±0.000802	0.009827±0.000320	0.000022±0.000004	48.011±0.692	0.998959±0.000159
Endpoint + Path	0.010090±0.000893	**0.009710±0.000350**	0.000023±0.000005	47.741±0.951	0.998891±0.000211
Endpoint + Composition	**0.009464±0.000790**	0.009760±0.000329	**0.000021±0.000004**	**48.241±0.836**	**0.999008±0.000167**
Endpoint + Path + Composition	0.010094±0.000852	0.009731±0.000329	0.000023±0.000005	47.719±0.954	0.998888±0.000207

Table 15: Experiment 4 direct–composed diagnostic at the held-out terminal frame using $\Delta = 4$ and temporal partition $(2, 2)$, summarized across 3 independently trained seeds.

Method	Direct FDE ↓	Rollout FDE ↓	CompErr ↓	GeoCompErr ↓	Render comp. MSE ↓
Endpoint	0.03461±0.00346	0.03495±0.00376	0.00867±0.00308	0.00802±0.00086	0.000011±0.000003
Endpoint + Path	0.03647±0.00401	0.03654±0.00420	0.00140±0.00020	0.00600±0.00105	**0.000006±0.000002**
Endpoint + Composition	**0.03360±0.00354**	**0.03368±0.00368**	0.00205±0.00047	0.00649±0.00133	0.000008±0.000003
Endpoint + Path + Composition	0.03654±0.00394	0.03654±0.00414	**0.00126±0.00019**	**0.00583±0.00105**	**0.000006±0.000002**

The same fixed RGB colors c_i and opacities α_i are used to render the source, predicted, and target Gaussian states. Consequently, image-space differences arise from predicted center and covariance evolution rather than an independently learned appearance model.

F.5.4 Held-out next-frame results

For the terminal one-step prediction we report Gaussian FDE, covariance-coordinate error, image MSE, PSNR, and SSIM. LPIPS was not enabled in the reported run and is therefore not reported.

The complete learned-model results are

The deterministic causal baselines obtain

$$\begin{aligned} \text{Last-frame:} \quad & \text{FDE} = 0.039723, \qquad \text{PSNR} = 35.304\,\text{dB}, \text{SSIM} = 0.984595, \\ \text{Constant velocity:} \quad & \text{FDE} = 0.002054, \qquad \text{PSNR} = 60.477\,\text{dB}, \text{SSIM} = 0.999938. \end{aligned}$$

The corresponding standard deviations across the 20 test trajectories are 0.012028, $3.237\,\text{dB}$, and 0.008493 for Last-frame, and 0.000740, $3.207\,\text{dB}$, and 0.000047 for constant velocity.

Endpoint + Composition is therefore the strongest learned configuration for held-out one-step center prediction and rendering, but causal constant velocity remains substantially stronger on this deliberately smooth short-horizon task. The experiment is consequently interpreted as evidence that learned Gaussian predictive geometry remains renderable at an unseen future endpoint, not as evidence of superiority to analytic local-motion extrapolation.

F.5.5 Held-out temporal-composition diagnostic

The $\Delta = 4$ diagnostic predicts the same held-out terminal state directly and through the temporal decomposition $(2, 2)$. The complete results are reported in table 15.

Relative to Endpoint, Endpoint + Composition reduces CompErr by 76.4% and GeoCompErr by 19.0%, while also slightly improving direct and rollout target-relative FDE at this horizon. The full Endpoint + Path + Composition model gives the smallest latent and Gaussian-level direct–composed discrepancies, reducing them by 85.4% and 27.3%, respectively, although its target-relative FDE is slightly higher. This reproduces the accuracy–consistency distinction observed in Experiment 1.

The direct and composed rendered images exhibit the same qualitative behavior. Their image-space MSE decreases from approximately 1.1×10^{-5} for Endpoint to 0.8×10^{-5} for Endpoint + Composition and approximately 0.6×10^{-5} for the path-supervised variants.

Interpretation. Experiment 4 closes the controlled chain from predictive latent state to explicit Gaussian geometry and finally to a rendered future observation. It does not establish general video-generation quality: appearance is fixed, the camera is fixed, the scenes are synthetic, and constant-velocity extrapolation is exceptionally strong at the one-step horizon. Its contribution is narrower and complementary to Experiments 1–3: the Gaussian state predicted by the non-reconstructive JEPA

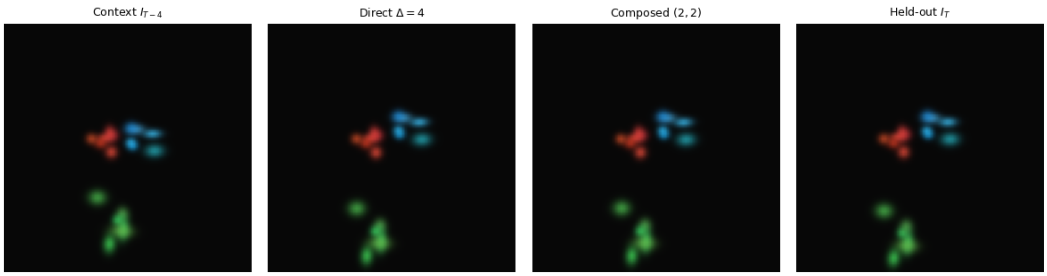


Figure 6: Held-out $\Delta = 4$ temporal-composition example for Endpoint + Composition (seed 0, test trajectory 0). The same terminal observation I_T is predicted directly from $T - 4$ and through the composed route $(2, 2)$. The direct and composed Gaussian predictions produce visually similar renderings while remaining close to the held-out target. This example was fixed by the pre-specified qualitative configuration rather than selected according to test performance.

objective remains usable by the standard rendering interface, and the temporal-consistency behavior of the latent and geometric predictions remains observable at a held-out rendered endpoint.

G Additional Discussion and Limitations

Scope of the current evidence. The experiments in this paper are controlled mechanism-level tests rather than a comprehensive evaluation of dynamic-scene world modelling. Experiments 1 and 2 operate directly on procedurally generated Gaussian states, providing exact future geometry, motion-group transformations, and temporal decompositions while bypassing the difficulty of constructing Gaussian scenes from real observations. Experiment 3 similarly isolates correspondence using synthetic identity and topology perturbations with known lineage. The reported results therefore support the proposed temporal-composition, geometry-composition, and correspondence mechanisms, but do not establish large-scale superiority over reconstruction-based Gaussian world models or video-space predictive models.

Why latent prediction rather than full Gaussian generation? A future scene contains both predictable structure and details that may be ambiguous, unobserved, or irrelevant to future dynamics. A reconstruction objective can allocate substantial capacity to reproducing precise appearance or representation-specific details even when they are not necessary for prediction. 4DGS-JEPA instead predicts future representations and uses the selective geometry decoder $Q_\psi^{\mathcal{G}}$ in section 3.4 to retain explicit geometric grounding.

This design does not imply that complete future Gaussian generation is undesirable. Rather, latent prediction and complete generation answer different questions. A generative Gaussian world model aims to produce a complete future scene, whereas 4DGS-JEPA learns a predictive state optimized to retain information relevant to future evolution.

What temporal composition does and does not guarantee. The temporal-composition loss $\mathcal{L}_{\text{comp}}$ in eq. (8) constrains alternative temporal routes to produce compatible latent states, while the geometry-composition loss $\mathcal{L}_{\text{gcomp}}$ in eq. (11) applies the analogous principle to structured decoded motion.

Neither objective establishes future correctness by itself. Two prediction routes may agree while both are incorrect. Target-relative supervision through $\mathcal{L}_{\text{end}}$, $\mathcal{L}_{\text{path}}$, and $\mathcal{L}_{\text{geo}}$ in eqs. (6), (7) and (39) remains necessary. Experiments 1 and 2 therefore evaluate composition jointly with target-relative prediction accuracy.

Temporal composition also does not imply an exact semigroup law for arbitrary dynamic scenes. Actions, absolute time, contacts, topology changes, and other non-autonomous effects may alter the

relevant transition. The appropriate requirement is consistency between admissible chronological decompositions of the same evolution, as formalized by the controlled composition law in eq. (27).

Dependence on the Gaussian front-end. The general formulation assumes that a sufficiently stable Gaussian representation can be constructed causally from observations through the online builder B_{on} in eq. (17). Errors in calibration, depth estimation, reconstruction, segmentation, motion grouping, tracking, or Gaussian initialization may therefore propagate into the predictive representation.

The present controlled experiments intentionally remove this source of error by supplying Gaussian states directly. Consequently, they test the predictive model independently of reconstruction quality but do not establish robustness to errors produced by a real online Gaussian front-end. Evaluation on causally reconstructed real dynamic scenes is required to characterize this interaction.

Deterministic future prediction. The transition Φ_ϕ in eq. (21) produces a single predictive future for a fixed context, horizon, and action sequence. This is appropriate for the approximately deterministic controlled trajectories evaluated here, but many real environments admit multiple plausible futures because of unobserved forces, agent decisions, or stochastic interactions.

In such settings, a deterministic transition may average incompatible outcomes or select one trajectory without representing its uncertainty. Extending the Gaussian-world transition to probabilistic or multimodal latent prediction is therefore a natural extension of the present formulation.

Gaussian identity and topology change. Persistent canonical identity provides reliable correspondence when the same canonical slot can be tracked through time. Experiment 3 shows, however, that raw array position is not a valid substitute for persistent canonical identity and that split, merge, birth, and death events require a more flexible correspondence mechanism.

The residual OT formulation in eq. (14) provides this fallback, but correspondence can remain intrinsically ambiguous when several primitives have nearly identical geometry, motion-group membership, and semantic features. Furthermore, the prefix-anchored canonical representation described in section 4 contains a finite pool of reserved inactive slots. Sufficiently large amounts of newly represented geometry could exhaust this capacity and require expansion or reconstruction of the canonical bank.

The present experiments also do not evaluate residual OT inside end-to-end predictive training: Experiments 1 and 2 use fixed persistent correspondence, whereas Experiment 3 studies correspondence independently of model optimization. The interaction between learned predictive states and residual correspondence under real topology-changing dynamics therefore remains to be evaluated.

Non-uniqueness of structured motion decomposition. The selective decoder represents future geometry through a combination of motion-group $\mathrm{SE}(3)$ transformations and Gaussian-specific residual deformation. This decomposition is generally not identifiable from endpoint geometry alone: different combinations of group motion and local residual deformation can produce similar final Gaussian states.

Experiment 2 demonstrates that the motion-group component of $\mathcal{L}_{\mathrm{gcomp}}$ substantially reduces path dependence in the predicted transformations, while the Gaussian-level component improves primitive-level consistency. These constraints encourage a common structured decomposition across temporal paths, but they do not imply that the resulting decomposition is uniquely or physically correct.

Structural priors. The rigidity, local smoothness, and static-region terms in eqs. (44) to (46) encode useful geometric biases but are not universal physical laws. Strong articulation, tearing, fluids, granular motion, abrupt contacts, or major topology changes may violate them. Such terms should therefore be interpreted as optional soft priors derived from causal evidence rather than hard physical constraints.

The present four experiments do not separately evaluate these structural regularizers. Their empirical contribution should therefore not be inferred from the current results.

Appearance and rendering. Experiments 1 and 2 deliberately isolate Gaussian geometry and contain no camera projection, rasterization, RGB supervision, PSNR, or perceptual image metrics.

Experiment 4 adds a controlled rendering evaluation and shows that predicted Gaussian geometry can be converted into a held-out future observation without introducing an RGB reconstruction loss into training.

This result does not yet establish learned future appearance prediction. In Experiment 4, Gaussian RGB colors and opacities are fixed over time and transported from the source state, the camera is fixed, and the scenes are procedurally generated. Consequently, its PSNR and SSIM primarily measure whether predicted future geometry projects to the correct image-space configuration. Evaluating view-dependent appearance evolution, illumination change, occlusion, and rendering from causally reconstructed real dynamic scenes remains future work.

Computational scaling. The controlled experiments use small Gaussian sets and known motion groups. Real Gaussian reconstructions may contain orders of magnitude more primitives. Although hierarchical grouping and persistent identity reduce the need for unrestricted primitive-wise comparison, attention over large Gaussian sets and residual OT matching can still become computationally expensive.

The present experiments do not contain a scaling study, so no empirical claim about large-scene computational efficiency is made. Practical large-scale implementations may require Gaussian-token subsampling, local or sparse attention, spatially restricted residual matching, or multiresolution grouping.

Action-conditioned dynamics and planning. The formulation in eqs. (21) and (27) permits future actions to condition the transition, but the four experiments reported here are passive forecasting, correspondence, and controlled rendering experiments. We therefore make no empirical claim about action-conditioned prediction, control, or planning in the present evaluation.

Testing whether temporal composition improves long-horizon action-conditioned rollout and model-based planning is an important extension, but it is separate from the mechanism-level evidence established in this paper.