QUAST: QUASICONFORMAL JACOBIAN-GUIDED SHAPE TRANSPORT FRAMEWORK FOR STABLE STREAMING 4D GAUSSIAN RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming 4D Gaussian reconstruction calls for local motion and Gaussian shape to evolve coherently across frames, yet existing residual-based update methods do not explicitly guarantee this. We present QuaST, a quasiconformal Jacobian-guided shape transport framework that models relative center motion and covariance evolution through a shared local deformation map. QuaST constructs a persistent motion-domain graph and estimates rank-aware local Jacobians from relative center motion. A quasiconformal-inspired prior constrains anisotropic distortion within local neighborhoods. A stop-gradient interface lets neighborhood motion guide covariance transport while routing the transport loss only to current Gaussian scale and rotation. Furthermore, temporal shape control limits excessive anisotropic changes of individual Gaussians across consecutive frames. QuaST integrates these constraints into optimization while retaining standard Gaussian rendering, without a learned deformation field or motion-driving anchors. QuaST is strictly causal: the update at frame t uses only the current training images and states stored from frames up to $t-1$, with no future-frame access. Experiments on DyNeRF and SelfCap show that QuaST improves novel-view reconstruction and maintains stable performance over extended captures of up to 1,200 frames.

1 INTRODUCTION

Streaming 4D reconstruction turns synchronized multiview video into a scene representation that can be updated incrementally and rendered from unseen viewpoints. Here, streaming is causal: the update at time t may use the synchronized views at t and states retained from earlier frames, but no observation from any frame $t' > t$. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) provides explicit, differentiable primitives with fast rasterization. Streaming methods such as 3DGStream (Sun et al., 2024) and QUEEN (Girish et al., 2024) exploit temporal redundancy by optimizing a detailed first frame and representing later frames as updates. The challenge is to preserve coherent geometry as these updates accumulate, while retaining novel-view fidelity and efficient rendering.

Such residual update schemes expose a structural ambiguity: methods like QUEEN learn center, rotation, and scale residuals without requiring them to explain the same deformation. Neighboring centers can move inconsistently, while individual Gaussians stretch or rotate in ways unsupported by that motion. Densification can improve the training-image fit without resolving either inconsistency. The desired stability is therefore not minimal change: a translating neighborhood should preserve Gaussian shape, a rotating one should rotate its Gaussian shapes, and genuine stretching should be expressed consistently by centers and covariances, without collapsing into implausible anisotropic distortion.

Our premise is that these two forms of inconsistency should be addressed through a shared *local deformation map*. A persistent neighborhood supplies relative center displacements from which a local Jacobian can be estimated. That Jacobian can both regularize neighborhood deformation and guide covariance evolution. It does not act as a motion anchor or prescribe a common translation: translation cancels in relative coordinates, and overlapping neighborhoods describe local deformation without imposing a global motion field.

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.

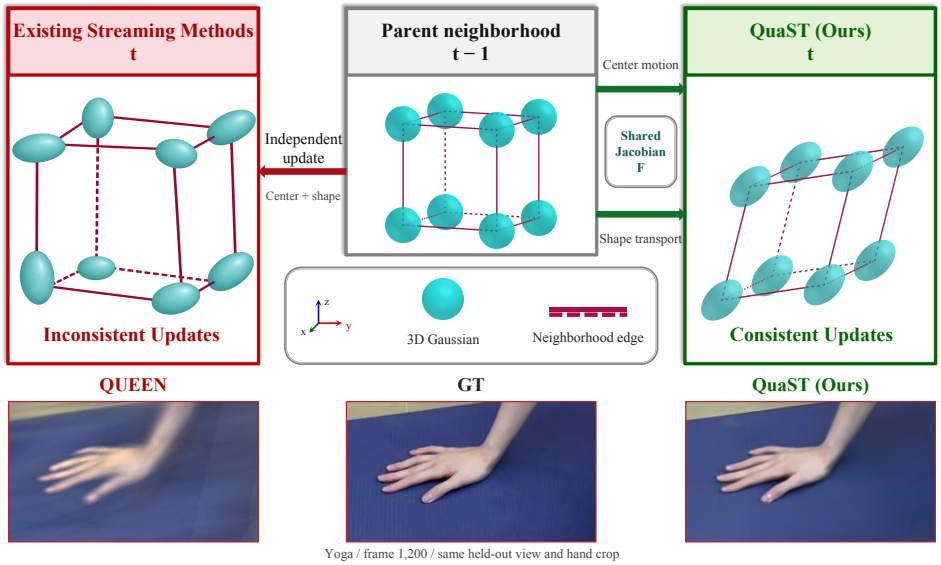


Figure 1: **Stability means compatible change, not frozen shape.** Top: a shared local Jacobian couples neighborhood motion and individual shape. Bottom: matched QUEEN, GT, and QuaST hand crops from Yoga at frame 1,200. QuaST preserves better-separated finger contours and a sharper palm boundary.

Three design considerations guide QuaST. Neighborhoods change under splitting, pruning, occlusion, and new visibility, so transient proximity does not establish shared motion. Surface-like sampling leaves normal deformation unobservable. Sharing a Jacobian between motion and shape objectives also requires explicit gradient routing. QuaST therefore estimates deformation on the observable neighborhood subspace and uses a detached copy for shape optimization.

We introduce *QuaST*, which couples local processing and individual-Gaussian processing through quasiconformal Jacobian-guided shape transport. A persistent graph in the RGB-derived dynamic domain rejects accidental proximity using geometric and co-motion tests. Each valid neighborhood yields a weighted least-squares Jacobian on its observable subspace. Local processing regularizes the local motion fit and penalizes anisotropic distortion. Individual processing transports the parent covariance with a stop-gradient Jacobian and complements this directional target with temporal quasiconformal shape control. The shared map links both levels, with transport gradients restricted to current covariance parameters.

Our contributions are:

1. We formulate motion–shape compatibility for streaming Gaussians: one local deformation map simultaneously governs both relative center motion and covariance evolution.
2. We develop a persistent motion-domain graph and a rank-aware Jacobian estimation scheme, enabling quasiconformal-inspired local regularization under changing topology and surface-like sampling.
3. We introduce gradient-decoupled shape transport and temporal shape control, using motion-derived targets to update current Gaussian scale and rotation.

2 Related Work

Dynamic neural and Gaussian representations. NeRF (Mildenhall et al., 2021) models static scenes with a continuous radiance field. Dynamic neural fields introduce canonical deformation (Pumarola et al., 2021) or factorized space–time planes (Fridovich-Keil et al., 2023; Cao & Johnson, 2023). Building on 3DGS (Kerbl et al., 2023), dynamic Gaussian methods use learned

deformation (Yang et al., 2024; Wu et al., 2024), persistent primitives (Luiten et al., 2024), or explicit spacetime Gaussian parameters (Li et al., 2024); SC-GS (Huang et al., 2024) instead uses sparse motion controls. Our graph measures overlapping local deformation rather than driving motion through controls.

Streaming free-viewpoint video. Pre-Gaussian streaming fields encode framewise model differences (Li et al., 2022a) or residual feature grids (Wang et al., 2023). Gaussian methods propagate cached transformations or attribute residuals (Sun et al., 2024; Girish et al., 2024), predict anchor-driven motion with keyframe refinement (Yan et al., 2025), and use hierarchical motion representations (Gao et al., 2024; Fu et al., 2026). Other designs propagate sparse keypoint motion (Chen et al., 2026), aggregate adjacent temporal features (Ke et al., 2026), or combine optical flow with per-Gaussian offsets and confidence (Lee et al., 2026). QuaST instead couples measured neighborhood deformation directly to Gaussian covariance evolution.

Local deformation priors. Embedded deformation blends graph-local affine transforms (Sumner et al., 2007); ARAP (Sorkine et al., 2007) penalizes departures from local rotations, and Dynamic 3D Gaussians (Luiten et al., 2024) applies local rigidity to persistent tracks. We fit a general observable linear map that exposes scale, shear, rank, and residual while bounding anisotropy. Covariance transport by a deformation gradient is established in continuum mechanics and PhysGaussian (Xie et al., 2024); here the gradient is estimated from reconstructed neighborhoods and used with explicit gradient decoupling.

Video segmentation and semantic motion priors. Frozen SAM 2 (Ravi et al., 2025) and DINOv2 (Oquab et al., 2023), combined with RGB motion evidence, produce dynamic, boundary, static, and uncertain regions. These masks define motion-domain support for graph construction; local deformation is estimated from reconstructed Gaussian neighborhoods. Uncertain Gaussians remain photometrically trainable outside structural losses.

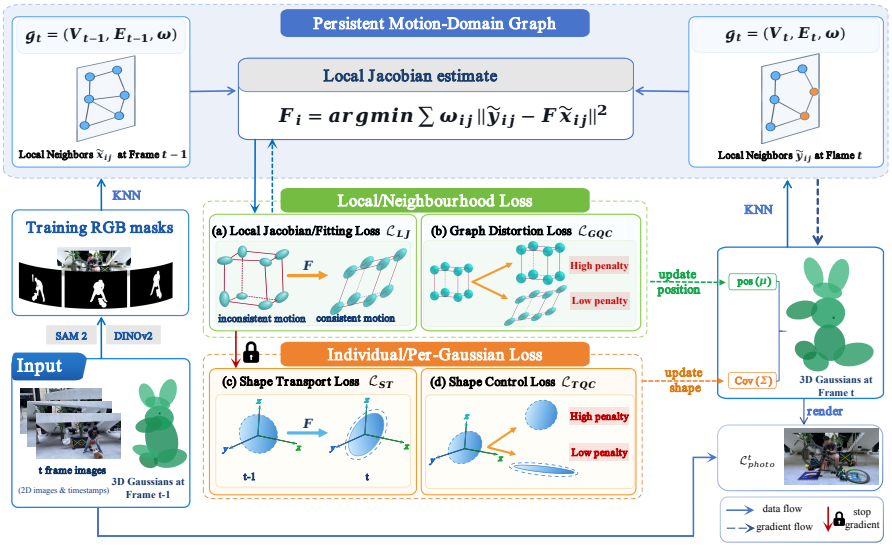


Figure 2: **Overview of a QuaST streaming update.** Causal RGB, SAM 2, and DINOv2 cues define motion-domain support for the persistent graph. Relative center displacements across frames estimate local Jacobians. Graph-QC regularizes directional stretch, Local-J fits and regularizes the local deformation map, and a detached Jacobian defines the covariance-transport target complemented by TQC. Joint photometric optimization updates Gaussian positions and covariances; solid/dashed arrows denote data/gradient flow, and the lock denotes stop-gradient.

3 METHOD

Figure 2 summarizes one streaming update. From frame-$(t-1)$ Gaussians and current images, causal RGB, SAM 2, and DINOv2 cues define motion-domain support for the persistent graph (Section 3.2). Relative center motion yields local Jacobians. Graph-QC regularizes directional stretch, while Local-J fits and regularizes the deformation map (Sections 3.3–3.4). A detached Jacobian supplies the covariance-transport target, complemented by temporal QC (Sections 3.5–3.6). Joint photometric optimization produces frame-t Gaussians, coupling neighborhood deformation and individual shape while routing shape-transport gradients only to current covariance parameters.

3.1 PRELIMINARIES: STREAMING GAUSSIANS

Given synchronized RGB images $\{I_{t,c}\}$ from calibrated training cameras $c \in \mathcal{C}_{\mathrm{tr}}$, we represent frame t with anisotropic Gaussians

$$G_i^t = \big(\boldsymbol{\mu}_i^t,\, \boldsymbol{q}_i^t,\, \boldsymbol{s}_i^t,\, \alpha_i^t,\, \boldsymbol{c}_i^t\big), \qquad \boldsymbol{\Sigma}_i^t = \boldsymbol{R}(\boldsymbol{q}_i^t)\,\mathrm{diag}\big((\boldsymbol{s}_i^t)^2\big)\boldsymbol{R}(\boldsymbol{q}_i^t)^\top, \tag{1}$$

Here, $\boldsymbol{\mu}_i^t$ is the center; $\boldsymbol{q}_i^t$ and $\boldsymbol{s}_i^t$ parameterize rotation and scale; and $\alpha_i^t, \boldsymbol{c}_i^t$ denote opacity and appearance.

Following QUEEN (Girish et al., 2024), multiview 3DGS fits the first frame; later Gaussians inherit decoded parent states through a map $p(i)$ and optimize attribute residuals. Independent attribute updates do not ensure consistency between center motion and covariance evolution. QuaST retains the renderer and adds geometric constraints only to predictive frames. Appendix A gives the full photometric objective, local differential derivation, and quasiconformal definitions.

3.2 PERSISTENT MOTION-DOMAIN GRAPH

We construct a persistent motion-domain graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t, w)$ for local deformation estimation, where $\mathcal{V}_t$ indexes dynamic Gaussian candidates, $\mathcal{E}_t$ contains accepted connections, and $w = \{w_{ij}\}$ denotes edge confidences. Its construction proceeds as follows:

Motion-domain candidates. Given the streaming Gaussian set $G^t = \{G_i^t\}_i$, RGB changes, forward-only SAM 2 propagation, and DINOv2 consistency provide dynamic masks from current and past training images. We take their union across consecutive frames to cover departure and arrival locations, then aggregate parent-Gaussian rendering contributions across training views to select

$$\mathcal{V}_t = \big\{ i \;\big|\; G_i^t \in G^t,\; v_i^t > 0,\; \rho_i^t \geq \tau_{\mathrm{dyn}},\; n_i^t \geq \min(m, v_i^t) \big\}. \tag{2}$$

Here, v_i^t counts visible training views, ρ_i^t is the fraction of total rendering influence within dynamic masks, and n_i^t counts dynamic-support votes. We use confidence threshold τ_{dyn} and $m = 2$, allowing single-view support when only one view is visible; Appendix A.4 details the voting rule. Static and uncertain Gaussians are excluded. For candidates with finite geometry and unique parent–child correspondences, we query K nearest dynamic neighbors at parent-frame centers to obtain $\mathcal{N}_K^t(i)$, excluding self-neighbors and newly added children.

Geometric and motion filtering. Covariance-aware contact tests parent-center separation relative to directional Gaussian support. Robust similarities at both endpoints assess motion agreement using median/MAD-standardized residuals, requiring a common cutoff, sufficient inliers, and observable rank. Refitting accepted neighborhoods supplies additional consistency confidence. These screening fits differ from the general Jacobian in Section 3.4.

Persistent edges and confidence. Let χ_{ij}^t combine lineage, contact, endpoint-motion, and refit tests, and q_{ij}^t denote temporal confirmation. The active edges are

$$\mathcal{E}_t = \big\{ (i,j) \in \mathcal{V}_t \times \mathcal{V}_t \;\big|\; j \in \mathcal{N}_K^t(i),\; i \in \mathcal{N}_K^t(j),\; \chi_{ij}^t q_{ij}^t = 1 \big\}. \tag{3}$$

Edges are confirmed over consecutive frames; missing edges retain history without exerting constraints. Weights $w_{ij} = h_{ij} \min(g_{i\to j}, g_{j\to i})\eta_{ij}$ combine contact, endpoint-motion, and refit confidences; ω_i averages them over each objective's valid support. Topology and weights remain detached and fixed between rebuilds. Appendix A.7 specifies these confidences and temporal updates.

3.3 Robust Graph Distortion

On the persistent graph, GQC regularizes differences in directional stretch within each neighborhood. For an active edge $(i,j) \in \mathcal{E}_t$, define its parent displacement $\boldsymbol{x}_{ij} = \boldsymbol{\mu}_{p(j)}^{t-1} - \boldsymbol{\mu}_{p(i)}^{t-1}$ and current displacement $\boldsymbol{y}_{ij} = \boldsymbol{\mu}_j^t - \boldsymbol{\mu}_i^t$. Let $\mathcal{N}_t(i) = \{j : (i,j) \in \mathcal{E}_t\}$ be the accepted neighbors of a valid graph center. We measure each edge's log stretch and the neighborhood's robust spread by

$$\begin{aligned} d_{ij} &= \log \frac{\|\boldsymbol{y}_{ij}\|_2 + \epsilon}{\|\boldsymbol{x}_{ij}\|_2 + \epsilon}, \\ D_i &= Q_{0.9}(\{d_{ij}\}; \{w_{ij}\}) - Q_{0.1}(\{d_{ij}\}; \{w_{ij}\})\,. \end{aligned} \tag{4}$$

Here, j ranges over $\mathcal{N}_t(i)$, and Q_q denotes the q-quantile weighted by the fixed edge confidences w_{ij}. Thus, D_i measures the spread of relative expansion across supported directions. The 90th–10th percentile range reduces sensitivity to isolated erroneous edges.

Given an allowed stretch-ratio parameter $K_0 \geq 1$, the objective is

$$\mathcal{L}_{\mathrm{GQC}} = \frac{\sum_i \omega_i \,\mathrm{ReLU}\left(D_i - \log K_0\right)^2}{\sum_i \omega_i + \epsilon}, \tag{5}$$

The sum covers valid graph centers; $\omega_i \geq 0$ is their confidence, and $\epsilon > 0$ provides numerical stabilization. Only spreads exceeding $\log K_0$ are penalized, leaving admissible local deformation unconstrained by this term.

In the ideal expression without stabilization, a common similarity transform $\boldsymbol{y}_{ij} = a\boldsymbol{R}\boldsymbol{x}_{ij}$, $a > 0$, gives $d_{ij} = \log a$ for all neighbors and hence $D_i = 0$. Relative edges cancel translation, while the spread is invariant to common rotation and scale. GQC therefore constrains directional inconsistency rather than absolute motion. It complements the subsequent local-Jacobian formulation, which estimates the explicit deformation map for shape transport.

3.4 Explicit Rank-Aware Local Jacobian

For each valid center, we compute weighted edge means $\bar{\boldsymbol{x}}_i$ and $\bar{\boldsymbol{y}}_i$ and define $\widetilde{\boldsymbol{x}}_{ij} = \boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i$ and $\widetilde{\boldsymbol{y}}_{ij} = \boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_i$. The local linearization solves

$$\boldsymbol{F}_i = \arg\min_{\boldsymbol{F}} \sum_j w_{ij} \left\|\widetilde{\boldsymbol{y}}_{ij} - \boldsymbol{F}\widetilde{\boldsymbol{x}}_{ij}\right\|_2^2 = \boldsymbol{B}_i \boldsymbol{C}_i^{+}, \tag{6}$$

with

$$\boldsymbol{C}_i = \sum_j w_{ij} \widetilde{\boldsymbol{x}}_{ij} \widetilde{\boldsymbol{x}}_{ij}^{\top}, \qquad \boldsymbol{B}_i = \sum_j w_{ij} \widetilde{\boldsymbol{y}}_{ij} \widetilde{\boldsymbol{x}}_{ij}^{\top}. \tag{7}$$

$\boldsymbol{C}_i$ is detached and cached, while $\boldsymbol{B}_i$ and $\boldsymbol{F}_i$ remain differentiable with respect to current centers.

Most visible Gaussians lie near a surface. We estimate deformation on the subspace supported by the parent neighborhood. Let $\gamma_1 \geq \gamma_2 \geq \gamma_3$ be the eigenvalues of $\boldsymbol{C}_i$. We retain directions satisfying $\gamma_k/\gamma_1 \geq \tau_{\mathrm{rank}}$ and compute the pseudoinverse $\boldsymbol{C}_i^{+}$. Rank-three neighborhoods use a full 3×3 Jacobian; rank-two neighborhoods use the observable tangent map; rank-one or insufficient neighborhoods do not receive a Jacobian loss. Rank selection reflects geometric observability.

Let $\sigma_{i,1}, \ldots, \sigma_{i,r}$ denote the positive singular values of $\boldsymbol{F}_i$ on the observable rank $r \in \{2,3\}$, and define $\overline{\log \sigma_i} = \frac{1}{r}\sum_{k=1}^{r} \log \sigma_{i,k}$. We use three interpretable terms:

$$\ell_{\mathrm{fit}}^i = \frac{\sum_j w_{ij} \|\widetilde{\boldsymbol{y}}_{ij} - \boldsymbol{F}_i \widetilde{\boldsymbol{x}}_{ij}\|_2^2}{\sum_j w_{ij} \|\widetilde{\boldsymbol{y}}_{ij}\|_2^2 + \epsilon}, \tag{8}$$

$$\ell_{\mathrm{aniso}}^i = \frac{1}{r}\sum_{k=1}^{r} \left(\log \sigma_{i,k} - \overline{\log \sigma_i}\right)^2, \qquad \ell_{\mathrm{scale}}^i = \left(\overline{\log \sigma_i}\right)^2. \tag{9}$$

The fitting term measures whether the neighborhood is well explained by a first-order deformation. The anisotropy term is a smooth, scale-invariant surrogate for the observable quasiconformal dilatation in Eq. (20): it ignores a common scale and penalizes disagreement among principal stretches. The

scale term supplies a separate near-isometric prior, with an independent coefficient for controlling uniform scale changes. The local Jacobian objective is

$$\mathcal{L}_{\mathrm{LJ}} = \frac{\sum_i \omega_i \left(\ell^i_{\mathrm{fit}} + \beta_{\mathrm{a}} \ell^i_{\mathrm{aniso}} + \beta_{\mathrm{s}} \ell^i_{\mathrm{scale}}\right)}{\sum_i \omega_i + \epsilon}. \tag{10}$$

Unlike ARAP, Eq. (10) does not first project every neighborhood to a rotation. It exposes the deformation actually supported by the data and then limits implausible anisotropy and scale drift.

3.5 Gradient-Decoupled Shape Transport

At a valid graph center, the neighborhood deformation in Eq. (6) also guides the evolution of Gaussian shape. Let $\widetilde{\boldsymbol{F}}_i = \boldsymbol{F}_i$ for rank-three neighborhoods; rank-two maps use the conservative completion in Appendix A.5. We transport the parent covariance as

$$\widehat{\boldsymbol{\Sigma}}^t_i = \bar{\boldsymbol{F}}_i \boldsymbol{\Sigma}^{t-1}_{p(i)} \bar{\boldsymbol{F}}_i^{\top}, \qquad \bar{\boldsymbol{F}}_i = \mathrm{sg}\left(\widetilde{\boldsymbol{F}}_i\right), \tag{11}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The affine-invariant distance on $\mathbb{S}^3_{++}$ compares the current covariance with this target:

$$\mathcal{L}_{\mathrm{ST}} = \frac{1}{3} \frac{\sum_i \omega_i \left\| \log\left[(\widehat{\boldsymbol{\Sigma}}^t_i)^{-\frac{1}{2}} \boldsymbol{\Sigma}^t_i (\widehat{\boldsymbol{\Sigma}}^t_i)^{-\frac{1}{2}} \right] \right\|_F^2}{\sum_i \omega_i + \epsilon}. \tag{12}$$

This metric compares relative shape, jointly accounting for rotation and anisotropic scale. Eq. (10) updates centers through the differentiable Jacobian; Eq. (12) updates only current scale and rotation, treating the motion-derived target as fixed during backpropagation.

3.6 Per-Gaussian Temporal Shape Control

Neighborhood transport supplies motion-conditioned directional targets at valid graph centers. Temporal shape control complements these targets for eligible dynamic Gaussians with valid parent correspondences. Define the Gaussian shape factor

$$\boldsymbol{A}^t_i = \boldsymbol{R}(\boldsymbol{q}^t_i)\,\mathrm{diag}(\boldsymbol{s}^t_i), \qquad \boldsymbol{J}^t_{i,\mathrm{shape}} = \boldsymbol{A}^t_i (\boldsymbol{A}^{t-1}_{p(i)})^{-1}. \tag{13}$$

Because the left and right rotations are orthogonal, the singular values of $\boldsymbol{J}^t_{i,\mathrm{shape}}$ equal the three parent-to-child axis-scale ratios. Let $\mathcal{P}_t$ denote the set of eligible dynamic Gaussians with valid parent correspondences. For each $i \in \mathcal{P}_t$, let $\Delta \log s_{i,k} = \log s^t_{i,k} - \log s^{t-1}_{p(i),k}$. We therefore compute its log condition number exactly from log-scale differences, without a matrix SVD:

$$\kappa^{\log}_i = \max_k \Delta \log s_{i,k} - \min_k \Delta \log s_{i,k}, \qquad \mathcal{L}_{\mathrm{TQC}} = \frac{1}{|\mathcal{P}_t|} \sum_{i \in \mathcal{P}_t} \left[\mathrm{ReLU}(\kappa^{\log}_i - \log K_{\mathrm{T}}) \right]^2. \tag{14}$$

This term penalizes temporal anisotropy beyond the allowed ratio $K_{\mathrm{T}} \geq 1$. TQC regulates parent-to-child axis-scale changes, while local Jacobian transport supplies a neighborhood-motion-conditioned directional target.

3.7 Training Objective and Streaming Schedule

For a predictive frame, the proposed temporal objective is

$$\mathcal{L}^t = \mathcal{L}^t_{\mathrm{photo}} + \lambda_{\mathrm{GQC}} \mathcal{L}_{\mathrm{GQC}} + \lambda_{\mathrm{LJ}} \mathcal{L}_{\mathrm{LJ}} + \lambda_{\mathrm{ST}} \mathcal{L}_{\mathrm{ST}} + \lambda_{\mathrm{TQC}} \mathcal{L}_{\mathrm{TQC}}. \tag{15}$$

QuaST is strictly streaming: initialization uses the first frame, and each subsequent update uses only the current training images and historical states, including mask propagation and graph maintenance. After inheriting the decoded parent state, photometric warm-up precedes a gradual ramp of structural weights. Discrete graph refreshes alternate with continuous attribute optimization. Appendix A.6 details topology updates and implementation; rendering uses ordinary decoded Gaussians.

4 EXPERIMENTS

We evaluate QuaST on the first 300 synchronized frames of six DyNeRF/N3DV scenes (Li et al., 2022b) and on 600-, 900-, and 1,200-frame SelfCap prefixes (Xu et al., 2024). We compare with 3DGStream (Sun et al., 2024), HiCoM (Gao et al., 2024), QUEEN (Girish et al., 2024), and MoRGS (Lee et al., 2026), reporting peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) (Wang et al., 2004), and standard LPIPS-VGG (Zhang et al., 2018). PSNR measures pixelwise reconstruction fidelity from the mean squared error between rendered and reference images and is expressed in decibels (dB). SSIM compares local luminance, contrast, and structure. Higher PSNR and SSIM and lower LPIPS indicate better quality. All QuaST sequences are processed in strict temporal order without future-frame access.

Evaluation protocol. DyNeRF/N3DV is evaluated at half resolution using published baseline values. SelfCap baselines are reproduced under the fixed camera, resolution, and full-image evaluation protocol in Appendix B. For each $T \in \{600, 900, 1200\}$, metrics average frames 1–T and all reported prefixes use the same runs.

Implementation details. We implement QuaST on QUEEN, retaining its first-frame multiview initialization and predictive frame-wise optimization. All experiments use Adam on a single NVIDIA A100-SXM4 GPU with 80 GB memory and PyTorch 2.5.0+cu121. Gaussian learning rates and schedules follow QUEEN. We use spherical harmonics of degree 2, 15 epochs per predictive frame (one pass over all training cameras per epoch), and an SSIM weight of 0.2. Structural terms are disabled during first-frame fitting; Appendix A.9 lists the remaining settings.

4.1 DYNERF/N3DV RESULTS

Table 1: DyNeRF/N3DV PSNR averaged over 300 frames at half resolution. Baseline values follow the original papers; best and second-best results are shown in bold and underline.

Method	Coffee Martini	Cook Spinach	Cut Beef	Flame Salmon	Flame Steak	Sear Steak	Mean
3DGStream	27.75	33.31	33.21	28.42	**34.30**	33.01	31.67
HiCoM	28.04	32.45	32.72	28.37	32.87	32.57	31.17
QUEEN	<u>28.38</u>	<u>33.40</u>	**34.01**	<u>29.25</u>	<u>34.17</u>	<u>33.93</u>	<u>32.19</u>
QuaST	**28.97**	**33.48**	<u>33.85</u>	**29.43**	34.16	**34.12**	**32.33**

Preserving reconstruction fidelity. QuaST achieves the highest mean PSNR among the compared methods: 32.33 dB, exceeding QUEEN by 0.14 dB and 3DGStream/HiCoM by 0.66/1.16 dB (Table 1). It improves over QUEEN on four of six scenes, with the largest gain on Coffee Martini (+0.59 dB). The remaining differences are small: QuaST is 0.16 dB lower on Cut Beef and 0.01 dB lower on Flame Steak. The benefit is therefore scene-dependent rather than uniform. These short-stream results show competitive reconstruction fidelity, while the longer SelfCap sequences below examine quality retention under repeated incremental updates.

4.2 RECONSTRUCTION QUALITY OVER 600 FRAMES

Table 2: SelfCap reconstruction quality over 600 frames. All baselines are reproduced under the same camera, resolution, and evaluation protocol. Best and second-best results are shown in bold and underline.

	Bike			Corgi			Yoga		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
HiCoM	23.43	0.810	0.392	20.72	0.722	0.499	13.56	0.673	0.593
MoRGS	25.44	0.882	0.254	22.02	<u>0.816</u>	<u>0.342</u>	18.37	<u>0.835</u>	**0.345**
3DGStream	<u>25.77</u>	<u>0.886</u>	<u>0.251</u>	21.48	0.783	0.349	17.89	0.808	0.392
QUEEN	24.00	0.846	0.307	<u>26.13</u>	0.776	0.424	<u>19.47</u>	0.826	0.398
QuaST	**25.89**	**0.893**	**0.233**	**27.27**	**0.827**	**0.339**	**19.65**	**0.854**	<u>0.375</u>

Quantitative comparison. QuaST achieves the highest PSNR and SSIM on all three SelfCap scenes (Table 2). On Bike, its 0.12 dB margin over 3DGStream is accompanied by higher SSIM (0.893 versus 0.886) and lower LPIPS (0.233 versus 0.251). Corgi shows a larger PSNR advantage,

exceeding QUEEN and 3DGStream by 1.14/5.79 dB while improving both SSIM and LPIPS over QUEEN. On Yoga, QuaST reaches 19.65 dB with 0.854 SSIM, but MoRGS retains lower LPIPS (0.345 versus 0.375). Thus, the gain extends beyond pixelwise error, although the best perceptual score is not attained in every scene.

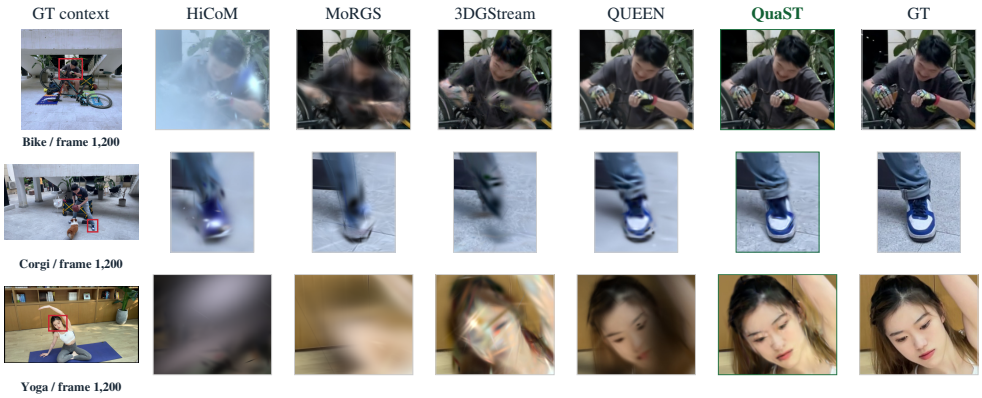


Figure 3: **Foreground detail at frame 1,200.** Red boxes locate the crop shared by all methods in each scene. Every reconstructed method corresponds to the reproduced runs in Tables 2–3, using the same camera, time, and image region.

Qualitative comparison. At frame 1,200, QuaST preserves sharper rider and glove contours on Bike, a more complete shoe outline on Corgi, and clearer facial detail on Yoga (Figure 3). The matched palm crops in Figure 1 show the same advantage for thin finger structures. These late-frame examples connect sequence-level scores to recognizable foreground boundaries and detail in the held-out views.

4.3 Long-Horizon Evaluation

Table 3: SelfCap reconstruction quality over 900- and 1,200-frame prefixes using the same reproduced runs and protocol as Table 2. Each value averages the complete prefix; best and second-best results are shown in bold and underline.

	Bike			Corgi			Yoga		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
900 frames									
HiCoM	20.19	0.779	0.430	21.40	0.721	0.507	12.63	0.630	0.634
MoRGS	25.34	0.880	0.256	22.33	<u>0.815</u>	<u>0.343</u>	18.35	<u>0.835</u>	**0.341**
3DGStream	<u>25.81</u>	<u>0.886</u>	<u>0.252</u>	21.43	0.782	0.351	18.12	0.813	0.386
QUEEN	22.41	0.832	0.316	<u>26.04</u>	0.776	0.425	<u>18.79</u>	0.816	0.405
QuaST	**25.91**	**0.893**	**0.233**	**27.23**	**0.827**	**0.335**	**19.63**	**0.855**	<u>0.367</u>
1,200 frames									
HiCoM	17.73	0.749	0.465	21.75	0.721	0.511	12.07	0.604	0.660
MoRGS	25.16	0.879	0.256	22.59	<u>0.816</u>	<u>0.341</u>	<u>18.25</u>	<u>0.832</u>	**0.344**
3DGStream	<u>25.79</u>	<u>0.886</u>	<u>0.252</u>	21.45	0.781	0.353	18.18	0.813	0.385
QUEEN	22.62	0.832	0.316	<u>25.92</u>	0.775	0.425	17.91	0.800	0.418
QuaST	**25.88**	**0.893**	**0.233**	**27.27**	**0.828**	**0.331**	**19.63**	**0.855**	<u>0.361</u>

Quality over extended streams. QuaST's prefix-averaged PSNR varies by only 0.03/0.04/0.02 dB on Bike/Corgi/Yoga as the evaluated horizon grows from 600 to 1,200 frames (Tables 2–3). Its margin over QUEEN on Yoga increases from 0.18 to 1.72 dB, with 1,200-frame SSIM improving from 0.800 to 0.855 and LPIPS from 0.418 to 0.361. Corgi shows a sustained advantage: the PSNR gap over QUEEN is 1.14/1.19/1.35 dB across the three prefixes, while QuaST's LPIPS decreases from 0.339 to 0.331. On Bike, 3DGStream is also stable, and QuaST retains a smaller 0.09 dB margin at 1,200 frames with better SSIM and LPIPS.

Stability throughout the stream. The unsmoothed curves in Figure 4 complement the cumulative averages by exposing when quality changes occur. On Bike, QuaST and 3DGStream remain close, whereas QUEEN shows a midstream decline followed by partial recovery. QuaST stays above

QUEEN and 3DGStream over most of Corgi and avoids QUEEN's pronounced late-stage drops on Yoga. The curves therefore support quality retention throughout the evaluated stream, rather than only at its endpoint. Together with the late-frame crops, they characterize the temporal persistence of reconstruction fidelity at both sequence and image-detail levels.

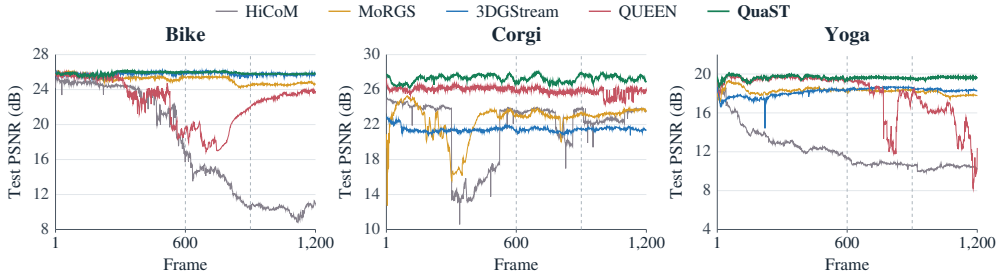


Figure 4: **Reconstruction quality throughout the stream.** Unsmoothed per-frame PSNR for QuaST, QUEEN, 3DGStream, MoRGS, and HiCoM, using the runs reported in Tables 2–3. All 1,200 measurements are retained; dashed lines mark frames 600 and 900.

4.4 Progressive Ablation of Structural Processing

Table 4: Progressive ablation at 1,200 frames. Each row cumulatively adds the indicated component; Full adds ST. N is the final Gaussian count, and Time is pure training time over frames 2–T (Appendix A.9).

	Yoga					Corgi				
Configuration	PSNR↑	SSIM↑	LPIPS↓	N (M)	Time (h)	PSNR↑	SSIM↑	LPIPS↓	N (M)	Time (h)
Base	17.91	0.800	0.418	1.228	9.65	25.92	0.775	0.425	1.414	6.49
+TQC	18.96	0.819	0.403	1.567	8.49	26.01	0.776	0.428	1.373	5.25
+Graph-QC	19.20	0.837	0.362	0.571	4.63	25.98	0.786	0.384	0.655	5.14
+Local-J	19.21	0.837	0.363	0.626	4.94	26.38	0.789	0.383	0.646	5.45
Full	19.63	0.855	0.361	0.840	5.30	27.27	0.828	0.331	0.825	5.91

Incremental quality gains. The progressive configurations reveal complementary effects (Table 4). TQC improves Yoga PSNR by 1.05 dB over Base. The graph-equipped stage improves SSIM and LPIPS in both scenes while reducing Gaussian count, although Corgi PSNR changes little. Local-J then adds 0.40 dB on Corgi. Adding ST to this configuration improves Yoga/Corgi PSNR by a further 0.42/0.89 dB and SSIM by 0.018/0.039. These conditional gains support adding a motion-conditioned covariance target to local regularization and temporal shape control.

Quality–cost trade-off. Full reaches 19.63/27.27 dB on Yoga/Corgi with 31.6%/41.7% fewer Gaussians than Base. Compared with +Local-J, it uses more Gaussians and 7.3%/8.4% more pure training time, but remains below Base in both count and time. The graph-equipped intermediate configurations offer more compact alternatives, while Full attains the strongest joint quality at 1,200 frames. The timings describe these recorded runs under the scope in Appendix A.9.

5 Conclusion

QuaST estimates a shared local Jacobian for Gaussian center motion and covariance evolution. Gradient-decoupled transport and temporal quasiconformal control improve strictly streaming reconstruction on DyNeRF and SelfCap, including 1,200-frame sequences, without suppressing genuine deformation.

Limitations. Temporal regularization cannot recover missing initial geometry, while occlusion, newly visible regions, and RGB segmentation errors can reduce graph reliability. Graph maintenance grows with Gaussian count and neighborhood size, and the local priors do not establish a globally quasiconformal homeomorphism. More efficient graph updates and confidence-aware scheduling remain future work.

AI USE STATEMENT

We used generative AI tools (large language models) for supporting tasks. These tools assisted with code implementation, including evaluation scripts and experiment orchestration, and with parts of data analysis. They also assisted with language polishing, translation, manuscript drafting and revision, and feedback on methodological and experimental explanations. The research question, method design, experimental protocol, and final conclusions were proposed and determined by the authors. All AI-assisted code, analyses, and text were reviewed by the authors, who take full responsibility for the content of this paper.

REFERENCES

Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 130–141. IEEE, 2023.

Jiacong Chen, Qingyu Mao, Youneng Bao, Xiandong Meng, Fanyang Meng, Ronggang Wang, and Yongsheng Liang. Motion matters: Compact gaussian streaming for free-viewpoint video reconstruction. *Advances in Neural Information Processing Systems*, 38:120385–120409, 2026.

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12479–12488. IEEE, 2023.

Jiaye Fu, Qiankun Gao, Chengxiang Wen, Yanmin Wu, Siwei Ma, Jiaqi Zhang, and Jian Zhang. Recon-gs: Continuum-preserved gaussian streaming for fast and compact reconstruction of dynamic scenes. *Advances in Neural Information Processing Systems*, 38:134751–134778, 2026.

Qiankun Gao, Jiarui Meng, Chengxiang Wen, Jie Chen, and Jian Zhang. Hicom: Hierarchical coherent motion for dynamic streamable scenes with 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:80609–80633, 2024.

Sharath Girish, Tianye Li, Amrita Mazumdar, Abhinav Shrivastava, David Luebke, and Shalini De Mello. Queen: Quantized efficient encoding of dynamic gaussians for streaming free-viewpoint videos. *Advances in Neural Information Processing Systems*, 37:43435–43467, 2024.

Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4220–4230. IEEE, 2024.

Zhihui Ke, Yvyang Liu, Xiaobo Zhou, and Tie Qiu. Streamstgs: Streaming spatial and temporal gaussian grids for real-time free-viewpoint video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5629–5638, 2026.

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

Wonjoon Lee, Sungmin Woo, Donghyeong Kim, Jungho Lee, Sangheon Park, and Sangyoun Lee. Morgs: Efficient per-gaussian motion reasoning for streamable dynamic 3d scenes. *arXiv preprint arXiv:2603.25042*, 2026.

Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. Streaming radiance fields for 3d video synthesis. *Advances in Neural Information Processing Systems*, 35:13485–13498, 2022a.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5511–5521. IEEE, 2022b.

Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8508–8520. IEEE, 2024.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pp. 800–809. IEEE, 2024.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10313–10322. IEEE, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Olga Sorkine, Marc Alexa, et al. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pp. 109–116, 2007.

Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pp. 80–es. 2007.

Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20675–20685. IEEE, 2024.

Jussi Väisälä. *Lectures on n-dimensional quasiconformal mappings*. Springer, 2006.

Liao Wang, Qiang Hu, Qihan He, Ziyu Wang, Jingyi Yu, Tinne Tuytelaars, Lan Xu, and Minye Wu. Neural residual radiance fields for streamably free-viewpoint videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 76–87. IEEE, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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. IEEE, 2024.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4389–4398. IEEE, 2024.

Zhen Xu, Yinghao Xu, Zhiyuan Yu, Sida Peng, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. Representing long volumetric video with temporal gaussian hierarchy. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024.

Jinbo Yan, Rui Peng, Zhiyan Wang, Luyang Tang, Jiayu Yang, Jie Liang, Jiahao Wu, and Ronggang Wang. Instant gaussian stream: Fast and generalizable streaming of dynamic scene reconstruction via gaussian splatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16520–16531. IEEE, 2025.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20331–20341. IEEE, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

A Additional Preliminaries and Local Differential Derivations

A.1 Photometric Optimization

The first frame is optimized with the standard multiview 3DGS objective and stored as the reference frame. For every $t > 1$, a parent map $p(i)$ associates a current Gaussian with its predecessor, and current attributes are initialized from the decoded parent plus learnable residuals as in QUEEN (Girish et al., 2024). The rendering objective is

$$\mathcal{L}^t_{\text{photo}} = (1 - \lambda_{\text{ssim}}) \, \|\widehat{I}_{t,c} - I_{t,c}\|_1 + \lambda_{\text{ssim}} \big(1 - \text{SSIM}(\widehat{I}_{t,c}, I_{t,c})\big), \tag{16}$$

for a sampled training camera c. Our method adds structure only to predictive frames; it does not change the renderer, appearance model, or the definition of Eq. (16).

A.2 Local Differential Motion–Shape Compatibility

Consider an unknown deformation $\boldsymbol{\varphi}_t$ that maps a parent-frame neighborhood to the current frame. Around the parent of Gaussian i, its first-order behavior is

$$\boldsymbol{\varphi}_t(\boldsymbol{\mu}^{t-1}_{p(i)} + \delta\boldsymbol{x}) \approx \boldsymbol{\varphi}_t(\boldsymbol{\mu}^{t-1}_{p(i)}) + \boldsymbol{F}^\star_i \delta\boldsymbol{x}, \qquad \boldsymbol{F}^\star_i = \nabla\boldsymbol{\varphi}_t(\boldsymbol{\mu}^{t-1}_{p(i)}). \tag{17}$$

The translation is contained in the first term and therefore disappears when relative offsets are compared. The Jacobian $\boldsymbol{F}^\star_i$ records the local rotation, scale, and shear. In particular, a translating neighborhood has $\boldsymbol{F}^\star_i = \boldsymbol{I}$ and should not induce any change of Gaussian shape, while a rigidly rotating neighborhood has $\boldsymbol{F}^\star_i = \boldsymbol{R}$.

Let $\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{\mu}^{t-1}_{p(i)}, \boldsymbol{\Sigma}^{t-1}_{p(i)}\right)$ represent the parent Gaussian. Under the first-order approximation in Eq. (17),

$$\boldsymbol{\varphi}_t(\boldsymbol{X}) \approx \boldsymbol{F}^\star_i \left(\boldsymbol{X} - \boldsymbol{\mu}^{t-1}_{p(i)}\right) + \boldsymbol{\varphi}_t\left(\boldsymbol{\mu}^{t-1}_{p(i)}\right). \tag{18}$$

Since an affine transformation $\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{X} + \boldsymbol{b}$ satisfies $\text{Cov}(\boldsymbol{Y}) = \boldsymbol{F}\,\text{Cov}(\boldsymbol{X})\boldsymbol{F}^\top$, the covariance therefore transforms as

$$\boldsymbol{\Sigma}^t_i \approx \boldsymbol{F}^\star_i \boldsymbol{\Sigma}^{t-1}_{p(i)} (\boldsymbol{F}^\star_i)^\top. \tag{19}$$

The relation is exact for an affine deformation and follows from the local first-order approximation for a general $\boldsymbol{\varphi}_t$.

Equations (17) and (19) expose the compatibility condition used in this work: relative center motion and Gaussian shape evolution should describe the same local deformation. They do not impose a global deformation field or require neighboring Gaussians to share one translation. QuaST estimates this map through persistent, rank-aware neighborhoods and separates the motion and shape gradients.

A.3 Quasiconformal Distortion on Observable Neighborhoods

Classical quasiconformal analysis bounds the directional distortion of a locally invertible, orientation-preserving map (Väisälä, 2006). For a full-rank differential $\boldsymbol{F}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_d > 0$, its linear dilatation is

$$K(\boldsymbol{F}) = \frac{\sigma_1}{\sigma_d}. \tag{20}$$

A conformal local map has $K = 1$: all directions undergo the same scale, possibly together with a rotation. Thus, quasiconformal distortion is invariant to translation and does not by itself penalize uniform scaling. Rigid motion is the special case $\boldsymbol{F} = \boldsymbol{R}$, whereas $\boldsymbol{F} = a\boldsymbol{R}$ remains conformal for any $a > 0$. This distinction is important in our formulation: anisotropic distortion and absolute scale drift are controlled by separate terms.

Gaussian neighborhoods observed by cameras often sample a two-dimensional surface rather than a three-dimensional volume. Let r be the rank supported by the parent neighborhood and let $\sigma_{i,1} \geq \cdots \geq \sigma_{i,r} > 0$ be the singular values of the local map restricted to that observable subspace. We use the rank-aware quantity $K_r(\boldsymbol{F}_i) = \sigma_{i,1}/\sigma_{i,r}$; rank-two neighborhoods constrain tangent deformation, while rank-one neighborhoods provide insufficient directional evidence and are excluded. Our quasiconformal-inspired regularizer controls dilatation on each observable local subspace. Sections 3.2–3.4 give the graph, estimator, and differentiable surrogate implementing this principle.

A.4 Motion-Domain Voting from Training Images

For training camera $c \in \mathcal{C}_{\mathrm{tr}}$ and frame s, define the dynamic mask $D_{s,c} = M^{\mathrm{core}}_{s,c} \cup M^{\mathrm{bnd}}_{s,c}$ and its temporal union $\widetilde{D}_{t,c} = D_{t-1,c} \cup D_{t,c}$. Static evidence requires agreement between both available frames outside this union; other pixels remain uncertain. Rendering the previous-frame parent gives nonnegative influence $a^t_{i,c}$ within the dynamic mask and total influence $b^t_{i,c}$. The visible-view count and dynamic-support vote count are

$$v^t_i = \sum_{c \in \mathcal{C}_{\mathrm{tr}}} \mathbf{1}[b^t_{i,c} > 0], \qquad n^t_i = \sum_{c \in \mathcal{C}_{\mathrm{tr}}} \mathbf{1}\left[\frac{a^t_{i,c}}{\max(b^t_{i,c}, \epsilon)} \geq \frac{1}{2}\right]. \tag{21}$$

The dynamic confidence and graph vertex set are

$$\rho^t_i = \frac{\sum_{c \in \mathcal{C}_{\mathrm{tr}}} a^t_{i,c}}{\max(\sum_{c \in \mathcal{C}_{\mathrm{tr}}} b^t_{i,c}, \epsilon)}, \qquad \mathcal{V}_t = \{i : v^t_i > 0,\ \rho^t_i \geq \tau_{\mathrm{dyn}},\ n^t_i \geq \min(m, v^t_i)\}. \tag{22}$$

Here τ_{dyn} controls dynamic confidence and $m = 2$. We require two dynamic-support votes when at least two training cameras are visible, and one vote when only one is visible. In both cases $\rho^t_i \geq \tau_{\mathrm{dyn}}$ must hold; zero-visibility Gaussians are rejected. Static and uncertain Gaussians are excluded from the structural graph. Dynamic masks use current and historical training RGB only.

A.5 Rank-Two Completion for Covariance Transport

For rank-three neighborhoods, the transport map is $\widetilde{\boldsymbol{F}}_i = \boldsymbol{F}_i = \boldsymbol{B}_i \boldsymbol{C}_i^{+}$. For a rank-two neighborhood, let $\boldsymbol{U}_{p,i} \in \mathbb{R}^{3\times 2}$ contain the two retained eigen-directions of $\boldsymbol{C}_i$, and let $\boldsymbol{n}_{p,i}$ be a unit normal to this subspace. Because $\boldsymbol{C}_i^{+}\boldsymbol{n}_{p,i} = \mathbf{0}$, the estimated Jacobian satisfies $\boldsymbol{F}_i\boldsymbol{n}_{p,i} = \mathbf{0}$; directly using $\boldsymbol{F}_i$ for covariance transport would therefore interpret the unobserved normal deformation as zero.

We extract the observed tangent deformation through the thin SVD

$$\boldsymbol{F}_i\boldsymbol{U}_{p,i} = \boldsymbol{U}_{t,i}\boldsymbol{S}_i\boldsymbol{V}_i^{\top}, \tag{23}$$

where $\boldsymbol{U}_{t,i} \in \mathbb{R}^{3\times 2}$ is an orthonormal basis of the mapped tangent plane. Its two-dimensional representation is

$$\boldsymbol{H}_i = \boldsymbol{U}_{t,i}^{\top}\boldsymbol{F}_i\boldsymbol{U}_{p,i} = \boldsymbol{S}_i\boldsymbol{V}_i^{\top}. \tag{24}$$

Thus, for any tangent coordinate $\boldsymbol{\xi}$, $\boldsymbol{U}_{p,i}\boldsymbol{\xi}$ is mapped to $\boldsymbol{U}_{t,i}\boldsymbol{H}_i\boldsymbol{\xi}$.

Let $\boldsymbol{n}_{t,i}$ be the consistently oriented unit normal of the mapped tangent plane. Since normal deformation is unobserved, we conservatively set $\boldsymbol{n}_{p,i} \mapsto \boldsymbol{n}_{t,i}$, preserving unit normal scale. Combining the tangent and normal actions gives

$$\widetilde{\boldsymbol{F}}_i = \boldsymbol{U}_{t,i}\boldsymbol{H}_i\boldsymbol{U}_{p,i}^{\top} + \boldsymbol{n}_{t,i}\boldsymbol{n}_{p,i}^{\top}. \tag{25}$$

The completion preserves the observed rank-two deformation without interpreting the unobserved normal direction as zero deformation. Rank-one and insufficient neighborhoods do not supply a transport teacher. The completed map is detached before constructing the covariance target in Eq. (11). Eigenvalues in the covariance distance are clamped for numerical positive-definiteness, without changing the intended relative-shape comparison.

A.6 Streaming Schedule and Topology Updates

Each predictive frame starts with photometric warm-up, followed by a linear ramp of the structural weights over 20% of its optimization iterations. The warm-up permits image evidence to align motion before structural regularization becomes fully active. Graph topology is refreshed periodically and whenever the topology version changes; graph decisions and weights are detached and fixed between refreshes. All temporal checks compare already available parent and current states. No later frame is required by mask propagation, neighborhood confirmation, or attribute optimization.

Clone, split, and prune operations remain active for dynamic and uncertain Gaussians. A child inherits parent-identity metadata and accumulates confirmation history before entering a structural neighborhood. Soft-static updates exclude confident static Gaussians from the graph but permit photometric correction. Hard-static updates additionally disable their geometric residual and densification. These update policies are separate from the Jacobian and covariance objectives.

The GPU implementation shares candidate neighborhoods across graph and codec bookkeeping, and batches contact tests, co-motion filtering, small-matrix decompositions, and structural losses. Identity and lifecycle counters remain discrete state. Candidate storage and filtering cost $O(|\mathcal{V}_t|K)$ for at most K candidates per dynamic Gaussian, while each local solve is 3×3. Decoded output Gaussians are rendered with the unchanged rasterizer.

A.7 Graph Construction: Definitions and Implementation

We specify the screening and confidence functions used in Section 3.2, suppressing time superscripts where the parent/current distinction is explicit.

Candidate eligibility and contact. Let P_{ij} indicate dynamic-domain membership, finite geometry, and valid parent identities at both endpoints. Each parent must map to a unique current child that is not newly added. Candidate neighborhoods use parent-frame centers and exclude self-neighbors.

For a nondegenerate parent edge, let $\ell_{ij} = \|\boldsymbol{x}_{ij}\|_2$ and $\boldsymbol{u}_{ij} = \boldsymbol{x}_{ij}/\ell_{ij}$. The directional standard deviation and normalized contact distance are

$$r_{i\to j} = \sqrt{\boldsymbol{u}_{ij}^\top \boldsymbol{\Sigma}_{p(i)}^{t-1} \boldsymbol{u}_{ij}}, \qquad c_{ij} = \frac{\ell_{ij}}{\sigma_c(r_{i\to j} + r_{j\to i})}. \tag{26}$$

Here $\sigma_c > 0$ controls support expansion and $\tau_c > 0$ bounds normalized separation. The contact indicator and confidence are $C_{ij} = \mathbf{1}[c_{ij} \le \tau_c]$ and

$$h_{ij} = C_{ij} \exp\left[-\frac{1}{2}(c_{ij}/\tau_c)^2\right]. \tag{27}$$

Nonfinite or degenerate edges are rejected. This directional proximity test is a covariance-aware contact proxy.

Robust similarity estimator. Given a source neighborhood, its eligible edge mask, and nonnegative base weights v_{ij}, fit an orientation-preserving similarity to the uncentered center-relative edges:

$$(a_i, \boldsymbol{R}_i) = \arg \min_{a>0,\, \boldsymbol{R}\in SO(3)} \sum_j v_{ij} \|\boldsymbol{y}_{ij} - a\boldsymbol{R}\boldsymbol{x}_{ij}\|_2^2. \tag{28}$$

Weighted orthogonal Procrustes gives the proper rotation, followed by a weighted least-squares scale. The relative residual is

$$e_{i\to j} = \frac{\|\boldsymbol{y}_{ij} - a_i \boldsymbol{R}_i \boldsymbol{x}_{ij}\|_2}{\max(a_i \|\boldsymbol{x}_{ij}\|_2, \epsilon)}. \tag{29}$$

The threshold therefore measures relative fit error; translations cancel in the center-relative edges.

Let m_i be the median eligible residual and let $s_i^e = 1.4826\ \mathrm{median}_j\, |e_{i\to j} - m_i| + 10^{-3}$. Starting from v_{ij}, two iteratively reweighted least-squares updates use

$$v_{ij}^{\mathrm{next}} = v_{ij} \min\left(1, \frac{m_i + z_0 s_i^e}{\max(e_{i\to j}, \epsilon)}\right), \qquad z_0 = 3, \tag{30}$$

The right-hand v_{ij} is the fixed base weight; residual statistics are recomputed at each update. The resulting fit selects provisional inliers by $e_{i\to j} \le m_i + z_0 s_i^e$. We refit with their original base weights and recompute their residual statistics. The final score and confidence are

$$z_{i\to j} = \max\left(0, \frac{e_{i\to j} - m_i}{s_i^e}\right), \qquad G(z) = \mathbf{1}[z \le z_0] \exp\left[-\frac{1}{2}(z/z_0)^2\right]. \tag{31}$$

The scale floor handles vanishing median absolute deviation. A reliable fit has finite parameters, at least six final inliers, an inlier fraction of at least 0.75, and parent-edge rank at least two. Rank is measured using the provisional-inlier base-weighted Gram matrix, retaining eigenvalues above 0.05 of its largest eigenvalue. This screening test is distinct from the centered rank test in Local-J.

Reciprocal endpoint screening and structural refit. The initial fit uses contact weights $v_{ij} = h_{ij}$. Denote its reliability by b_i and confidence by $g_{i\to j} = G(z_{i\to j})$. With the mutual-neighbor indicator $M_{ij} = \mathbf{1}[j \in \mathcal{N}_K^t(i),\ i \in \mathcal{N}_K^t(j)]$, preliminary acceptance and weights are

$$A_{ij} = M_{ij} P_{ij} C_{ij} b_i b_j \mathbf{1}[z_{i\to j} \le z_0] \mathbf{1}[z_{j\to i} \le z_0], \qquad \kappa_{ij} = A_{ij} h_{ij} \min(g_{i\to j}, g_{j\to i}). \tag{32}$$

The score at endpoint j uses its independently fitted neighborhood; both endpoint scores use only parent/current states.

Refitting the accepted neighborhood with weights κ_{ij} gives source reliability $\widetilde{b}_i$ and confidence $\eta_{ij} = G(\widetilde{z}_{ij})$, where $\widetilde{z}_{ij}$ is its standardized residual. The compatibility indicator in Eq. (3) is

$$\chi_{ij}^{t} = P_{ij} C_{ij} b_i b_j \widetilde{b}_i \mathbf{1}[z_{i\to j} \le z_0] \mathbf{1}[z_{j\to i} \le z_0]. \tag{33}$$

Mutual KNN is imposed separately in Eq. (3). The final weight $w_{ij} = h_{ij} \min(g_{i\to j}, g_{j\to i}) \eta_{ij}$ includes consistency after restricting the neighborhood to accepted edges. Refined outliers have zero weight. Source-wise refits may give different weights for the two endpoint-centered neighborhoods.

Framewise persistence and valid support. Temporal confirmation follows geometric screening and refitting. A new edge requires three consecutive accepted frames. Confirmed edges retain their history for up to eight missing frames with zero current weight; longer absence retires that history. Parent–child remapping preserves edge identities only when both endpoints remain unambiguous.

Within a frame, graph rebuilds read the same preceding-frame history; only the final rebuild commits the next history. Thus confirmation counts video frames rather than optimization iterations, without accessing future observations.

Each consuming objective averages the stored confidences over its valid support $\mathcal{S}_i$:

$$\omega_i = \frac{\sum_{j\in\mathcal{S}_i} w_{ij}}{|\mathcal{S}_i|}, \tag{34}$$

Invalid centers receive zero confidence. For Graph-QC, $\mathcal{S}_i$ contains active slots with finite, nonzero current length, with sufficient positive-weight edges required for validity. Local-J and ST use finite parent edges with valid indices and positive finite weights, requiring at least six positive-weight neighbors and rank at least two. Stored centers failing these loss-specific checks are excluded; graph membership alone does not imply validity. These objective-specific masks define the averaging support. Topology and confidences remain detached between rebuilds; structural losses differentiate Gaussian attributes, not screening decisions.

A.8 Visualization of Screened Motion-Domain Neighborhoods

Figure 5 shows the Corgi frame-900 checkpoint with 2,725 stored graph centers and 37,340 directed edges. Under $K = 32$, neighbor counts span 6–32 using the stored edge mask and 4–32 after excluding zero-weight edges; the latter have median 13 and are visualized. Fourteen centers have fewer than six positive-weight neighbors and are excluded from Local-J/ST; the others still require rank and other validity checks (Appendix A.7).

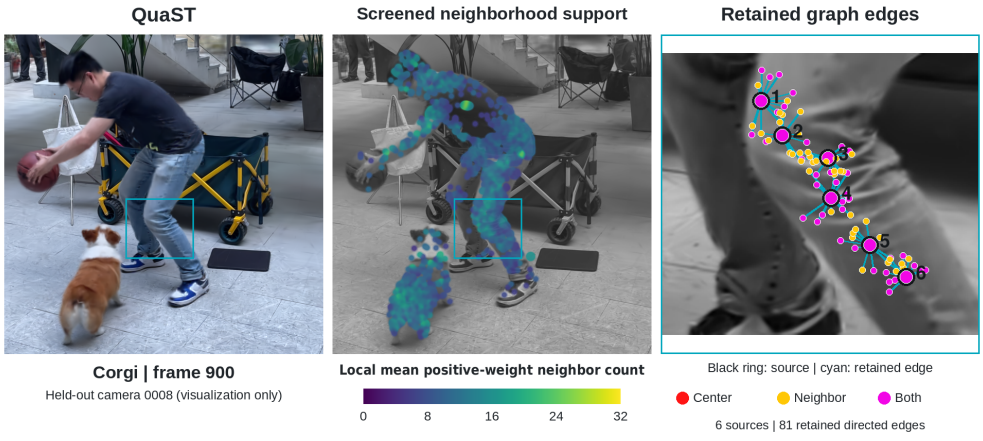


Figure 5: **Screened motion-domain neighborhoods on Corgi at frame 900.** The graph is constructed only from training views; held-out `camera_0008` is used solely to project stored nodes and edges for display. The middle panel shows the local mean positive-weight neighbor count of stored graph centers. The right panel displays six source nodes and their 81 retained directed edges from the same checkpoint.

A.9 Implementation Details

For temporal quasiconformal regularization, we use $\lambda_{\mathrm{TQC}} = 0.01$ and $K_T = 2.0$ on all datasets. On DyNeRF/N3DV, we set $\lambda_{\mathrm{GQC}} = 0.01$, $\lambda_{\mathrm{LJ}} = 0.01$, and $\lambda_{\mathrm{ST}} = 0.01$; on SelfCap, the corresponding weights are 0.025, 0.03, and 0.03. We use $K_0 = 1.5$ and $(\beta_a, \beta_s) = (1, 1)$ throughout. Graph-QC and Local-J use a warm-up fraction of 0.5 followed by a linear ramp of 0.2; individual temporal regularization uses a warm-up fraction of 0.2 and a ramp of 0.2.

We construct motion domains using frozen SAM 2 and DINOv2 features, which also initialize the learnable position gates. We use a mask threshold of 0.5 and dynamic/static consensus fractions of $0.55/0.70$. Dynamic support follows $n_i^t \geq \min(v_i^t, 2)$ with $v_i^t > 0$, including single-visible-view fallback and the dynamic-confidence threshold (Appendix A.4). The dynamic graph uses $K = 32$ nearest neighbors with at most 8,192 source Gaussians. We set $(\sigma_c, \tau_c) = (2.0, 1.5)$, the MAD threshold to 3.0, the minimum consensus ratio to 0.75, the minimum number of valid neighbors to 6, and the rank threshold to 0.05. Graph-QC uses weighted quantiles of 0.1 and 0.9. Edges are confirmed after three observations, retained for at most eight missing frames, and rebuilt every 32 optimization iterations. Shape transport is evaluated only at graph source centers passing Jacobian, parent–child correspondence, and shape-validity checks, with opacity above 0.01. Independently, the per-Gaussian TQC branch samples at most 65,536 eligible dynamic Gaussians with valid parent correspondences.

Densification is performed every two epochs with a gradient threshold of 1.25×10^{-3} and stops after 80% of predictive-frame optimization. DyNeRF/N3DV uses the hard-static policy, whereas SelfCap uses the soft-static policy. All timing measurements use the same single NVIDIA A100-SXM4 80 GB GPU. Pure training time is the GPU-synchronized elapsed time for predictive-frame training over frames 2–T; it excludes initialization, data loading, evaluation, and output storage.

B SelfCap Sequence and Evaluation Protocol

Table 5 specifies the three SelfCap sequences used in the quantitative comparisons. Source-frame indices are zero-based, whereas reconstruction steps are numbered from 1. Each horizon contains the first 600, 900, or 1,200 consecutive frames of the same sequence, without temporal subsampling. Camera IDs refer to physical cameras.

Table 5: SelfCap source intervals, camera splits, and evaluation dimensions (width × height). The start denotes the common source timeline; Corgi retains the synchronization offsets described below. Training-camera ranges are inclusive.

Scene	Start	Test camera	Training cameras	Evaluation
Bike	8,900	0010	0001–0009, 0011–0022 (21 cameras)	1079×1079
Corgi	5,203	0008	0001–0007, 0009–0024 (23 cameras)	1919×1079
Yoga	0	0012	0002–0011, 0013–0024 (22 cameras)	1920×1080

Temporal alignment. Bike uses source frames 8,900–9,499, 8,900–9,799, and 8,900–10,099 for the three horizons. The corresponding Yoga intervals are 0–599, 0–899, and 0–1,199. Corgi uses common-timeline intervals 5,203–5,802, 5,203–6,102, and 5,203–6,402. Its released clip begins at source frame 5,200, so the retained sequence begins at clip index 3. Camera-specific nearest-frame synchronization uses the supplied time offsets at 60 fps, with integer corrections between -3 and $+2$ frames; the held-out camera 0008 has zero integer correction. These camera-specific offsets align images to the same physical capture time. Bike and Yoga use direct frame indices. For every horizon $T \in \{600, 900, 1200\}$, the reported scores average all frames 1–T. Each prefix is evaluated as causal processing reaches it.

Camera membership. The test camera is excluded from photometric optimization at every frame. Yoga also excludes camera 0001 because of camera instability. Its retained cameras are therefore 0002–0024, with 0012 held out. Bike and Corgi retain all cameras other than their test camera for training.

Spatial preprocessing and evaluation region. Bike uses the calibrated, undistorted common valid image region at its native processed size of 1079×1079. Corgi is undistorted, cropped to the common valid 3839×2159 region, and downsampled by two to 1919×1079. Yoga uses the calibrated 1920×1080 images. Metrics evaluate the entire resulting image, including foreground and background; neither the SAM/DINO masks nor the red-box qualitative crops restrict the scoring region. PSNR and SSIM are averaged frame by frame, and LPIPS denotes standard VGG LPIPS with RGB normalized to $[-1, 1]$.

C Progressive Structural Configurations across Stream Lengths

Tables 6 and 7 extend the structural comparison to the 600-, 900-, and 1,200-frame prefixes of each configuration's continuous run. Each quality entry averages the complete indicated prefix using floating-point RGB and standard LPIPS-VGG; N counts Gaussians at the final frame. Time is pure training time over predictive frames 2–T, using the definition in Appendix A.9.

The row labels follow Table 4: Base is the QUEEN reference; +TQC adds individual temporal quasiconformal regularization; +Graph-QC adds persistent graph construction and graph quasiconformal regularization; +Local-J includes the local differential motion objective; and Full additionally includes shape transport. The Corgi and Yoga Base runs also provide the QUEEN results in Tables 2 and 3.

Table 6: Yoga cumulative configurations across 600, 900, and 1,200 frames. Metrics average frames 1–T; N is measured at T, and Time is pure training time over frames 2–T (Appendix A.9).

Configuration	PSNR↑	SSIM↑	LPIPS↓	N (M)	Time (h)
600 frames					
Base	19.47	0.826	0.398	0.961	2.27
+TQC	18.87	0.819	0.404	0.947	3.25
+Graph-QC	19.29	0.843	0.363	0.474	2.20
+Local-J	19.26	0.843	0.364	0.497	2.36
Full	19.65	0.854	0.375	0.409	2.40
900 frames					
Base	18.79	0.816	0.405	1.334	5.33
+TQC	18.92	0.820	0.404	1.260	5.63
+Graph-QC	19.27	0.841	0.361	0.515	3.38
+Local-J	19.23	0.840	0.362	0.555	3.62
Full	19.63	0.855	0.367	0.630	3.80
1,200 frames					
Base	17.91	0.800	0.418	1.228	9.65
+TQC	18.96	0.819	0.403	1.567	8.49
+Graph-QC	19.20	0.837	0.362	0.571	4.63
+Local-J	19.21	0.837	0.363	0.626	4.94
Full	19.63	0.855	0.361	0.840	5.30

Yoga. Full maintains 19.65/19.63/19.63 dB over the three prefixes while LPIPS improves from 0.375 to 0.361. At 1,200 frames it reaches 0.855 SSIM with 0.840M Gaussians, 31.6% fewer than Base; +Graph-QC provides a compact 0.571M-Gaussian operating point.

Table 7: Corgi cumulative configurations across 600, 900, and 1,200 frames. Metrics average frames 1–T; N is measured at T, and Time is pure training time over frames 2–T (Appendix A.9).

Configuration	PSNR↑	SSIM↑	LPIPS↓	N (M)	Time (h)
600 frames					
Base	26.13	0.776	0.424	0.913	2.71
+TQC	26.14	0.777	0.426	0.883	2.35
+Graph-QC	26.10	0.791	0.379	0.532	2.40
+Local-J	26.56	0.794	0.379	0.534	2.56
Full	27.27	0.827	0.339	0.617	2.82
900 frames					
Base	26.04	0.776	0.425	1.162	4.49
+TQC	26.06	0.775	0.428	1.127	3.71
+Graph-QC	25.99	0.788	0.382	0.592	3.72
+Local-J	26.42	0.791	0.381	0.588	3.95
Full	27.23	0.827	0.335	0.721	4.32
1,200 frames					
Base	25.92	0.775	0.425	1.414	6.49
+TQC	26.01	0.776	0.428	1.373	5.25
+Graph-QC	25.98	0.786	0.384	0.655	5.14
+Local-J	26.38	0.789	0.383	0.646	5.45
Full	27.27	0.828	0.331	0.825	5.91

Corgi. Full maintains 27.27/27.23/27.27 dB and improves LPIPS from 0.339 to 0.331. At 1,200 frames it reaches 0.828 SSIM with 0.825M Gaussians, 41.7% fewer than Base; the graph-equipped intermediate stages use approximately 0.65M Gaussians.