DDD3R: DIRECTIONAL DECOMPOSITION AND DAMPENING FOR RECURRENT 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

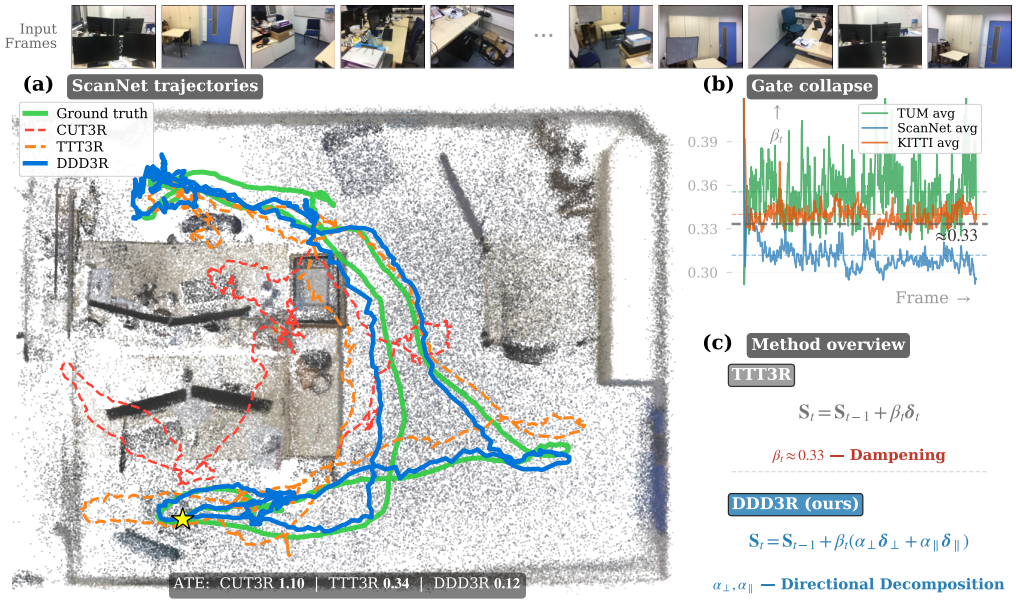

Figure 1: **DDD3R diagnoses and fixes state-update failures in recurrent 3D reconstruction.** **(a)** On ScanNet sequence, CUT3R and TTT3R drift severely (ATE 1.10 / 0.34) while DDD3R tracks the ground-truth trajectory closely (ATE 0.12). **(b)** The underlying cause: TTT3R's adaptive gate collapses to near-constant values on TUM, ScanNet, and KITTI ($\mu \in [0.31, 0.35]$), providing no real temporal adaptivity. **(c)** DDD3R replaces scalar dampening with directional decomposition, a training-free and parameter-free plug-in that attenuates drift ($\delta_\parallel$) while preserving novel content ($\delta_\perp$).

ABSTRACT

Recurrent 3D reconstruction models process video streams by iteratively updating a latent state, yet we find that this update process suffers from three compounding failures. Update magnitudes are systematically too large, causing progressive degradation over long sequences. All existing adaptive gating mechanisms collapse to near-constant values, providing no meaningful temporal adaptivity. And update vectors contain substantial directional redundancy, repeating historical drift rather than encoding new geometric information. These failures share one cause. Any scalar gate modulates the update uniformly across all directions in feature space, so it cannot distinguish redundant drift from novel content and reduces to blunt magnitude control. Motivated by this diagnosis, we propose **DDD3R** (**D**irectional **D**ecomposition and **D**ampening for Recurrent **3**D **R**econstruction), a training-free module that replaces scalar gating with directional decomposition. **DDD3R** projects each per-token state update onto a tracked drift direction and applies differentiated suppression, strongly attenuating drift while preserving orthogonal, novel information. This formulation strictly generalizes prior approaches. Constant dampening, temporal braking, and full directional decomposition all emerge as special cases of a single update rule parameterized by how much to differentiate drift from novelty. A drift-energy-adaptive extension parameterizes a continuous

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.

spectrum from full directional decomposition toward isotropic dampening that can be tuned per scene. Applied as a plug-in to existing models without retraining, a single fixed configuration ($\text{DDD3R}_{\text{brake}}$, the scalar relaxation of the full rule) reduces camera pose error by 62–68% on long indoor sequences (TUM, ScanNet at 1000 frames), and the directional-decomposition variant ($\text{DDD3R}_{\text{ortho}}$) yields a further $\sim$13% relative reduction over $\text{DDD3R}_{\text{brake}}$ on TUM where drift carries mostly harmful repetition. DDD3R also improves video depth estimation by 13–20% across three benchmarks (KITTI, Bonn, Sintel), while adding no learnable parameters and negligible computational overhead.

1 INTRODUCTION

Reconstructing 3D scenes from streaming video is a core capability for robotics, augmented reality, and autonomous navigation. Recent feed-forward models Wang et al. (2024); Leroy et al. (2024); Wang et al. (2025a) have achieved impressive results on multi-view 3D reconstruction, but their quadratic memory cost limits them to short, offline image collections. To handle continuous video streams, recurrent architectures Wang et al. (2025b); Chen et al. (2026) maintain a compact latent state that is updated frame-by-frame, enabling constant-memory inference at real-time speed. This recurrent formulation has rapidly become the dominant paradigm for streaming 3D reconstruction.

Yet the very mechanism that enables streaming, iterative state update, is also its weakest point. We present a systematic diagnosis revealing that state updates in existing recurrent 3D reconstruction models are fundamentally under-regulated, exposing three compounding failures. **(M1) Uncontrolled magnitude.** Updates are systematically too large, errors compounding 5.0–$8.6\times$ from short to long sequences, yet a constant scaling of 0.5 already cuts pose error by 32–65% (section 3.1). **(M2) Adaptive-gate collapse.** Learned gates such as TTT3R's sigmoid ($\sigma_{\text{temp}} \leq 0.03$) and TTSA3R's TAUM ($\sigma \approx 0.006$) saturate or normalise their input and degenerate into near-constants (section 3.2). **(M3) Directional redundancy.** Consecutive state deltas align strongly ($\cos \approx 0.62$–0.77, 40–60% of update energy along a persistent drift direction), and whether that drift carries harmful repetition or useful refinement is scene-dependent (section 3.3).

These three failures share a single root cause. Scalar gating operates only on the magnitude axis, leaving the directional structure of updates entirely uncontrolled. Constant dampening addresses Failure 1 but ignores Failures 2 and 3. Adaptive gates attempt to address Failure 2 but collapse back to constants. And directional redundancy is fundamentally inaccessible to any scalar mechanism. This motivates a shift from how much to update each token to which directions to preserve within each token's update. So we propose **DDD3R** (**D**irectional **D**ecomposition and **D**ampening for **R**ecurrent **3D R**econstruction), a training-free module that introduces directional control as a new axis orthogonal to existing scalar gating. DDD3R decomposes each per-token state update into a drift component (aligned with the historical update direction) and an orthogonal component (carrying novel geometric information), and applies differentiated suppression, strongly attenuating drift while preserving orthogonal content. This formulation yields a unified update rule,

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \beta_t(\alpha_\perp \boldsymbol{\delta}_t^\perp + \alpha_\parallel \boldsymbol{\delta}_t^\parallel), \tag{1}$$

that strictly generalizes all prior update strategies as special cases. Setting $\alpha_\perp = \alpha_\parallel$ recovers constant dampening, temporal braking is recovered as a scalar relaxation that collapses the directional axis, and the full directional decomposition is recovered when $\alpha_\perp \gg \alpha_\parallel$. A drift-energy-adaptive mechanism further allows the framework to self-adjust across scenes, automatically relaxing directional suppression when drift reflects useful refinement rather than harmful repetition.

Contributions. Our work makes three contributions.

(1) Diagnosis. We identify three compounding failures (M1–M3) in recurrent state updates and show that they are manifestations of a single limitation, namely that scalar gating cannot reach the directional axis.

(2) Unified framework. We propose DDD3R, a training-free directional-decomposition module whose update rule subsumes constant dampening, temporal braking, and full directional decomposition as boundary cases on one parameterized spectrum.

(3) Empirical results. Applied as a plug-in to CUT3R/TTT3R without retraining, the most consistent fixed configuration ($\text{DDD3R}_{\text{brake}}$) cuts camera pose error by 62–68% on long indoor sequences (TUM, ScanNet 1000f), and $\text{DDD3R}_{\text{ortho}}$ adds a further $\sim$13% relative reduction over $\text{DDD3R}_{\text{brake}}$ on TUM. We also improve video depth by 13–20% on KITTI, Bonn, and Sintel, and add zero parameters with negligible compute.

2 PRELIMINARIES

We build on TTT3R Chen et al. (2026), which extends CUT3R Wang et al. (2025b) with test-time learning for recurrent 3D reconstruction. The model maintains a persistent state $\mathbf{S}_{t-1} \in \mathbb{R}^{n\times c}$ consisting of n tokens with channel dimension c. At each timestep, an incoming observation is encoded into image tokens $\mathbf{X}_t \in \mathbb{R}^{(h\times w)\times c}$, and the state is updated via multi-head cross-attention. Queries are projected from the state and keys/values from the observation:

$$\mathbf{Q}_{\mathbf{S}_{t-1}} = \mathbf{S}_{t-1}\mathbf{W}_Q \in \mathbb{R}^{n\times c}, \qquad \mathbf{K}_{\mathbf{X}_t} = \mathbf{X}_t\mathbf{W}_K \in \mathbb{R}^{(h\times w)\times c}, \qquad \mathbf{V}_{\mathbf{X}_t} = \mathbf{X}_t\mathbf{W}_V \in \mathbb{R}^{(h\times w)\times c}. \tag{2}$$

CUT3R updates the state by full replacement with the cross-attention output:

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{\Delta}_t, \qquad \mathbf{\Delta}_t = \mathrm{softmax}\big(\mathbf{Q}_{\mathbf{S}_{t-1}}\mathbf{K}_{\mathbf{X}_t}^\top\big)\,\mathbf{V}_{\mathbf{X}_t}. \tag{3}$$

TTT3R reinterprets this as test-time learning and modulates the state update with an extra per-token learning rate $\beta_t \in \mathbb{R}^{n\times 1}$:

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \beta_t\mathbf{\Delta}_t, \qquad \beta_t = \mathrm{sigmoid}\big(\textstyle\sum_m \mathbf{Q}_{\mathbf{S}_{t-1}}\mathbf{K}_{\mathbf{X}_t}^\top\big), \tag{4}$$

where β_t is broadcast along the channel dimension and $\sum_m$ denotes the normalized summation over the spatial dimension $m = \{1,\dots,h\}\times\{1,\dots,w\}$. CUT3R thus corresponds to the special case $\beta_t \equiv 1.0$, while TTT3R derives β_t from cross-attention confidence and is intended to selectively suppress low-quality updates. For the analysis that follows we additionally write $\boldsymbol{\delta}_t \in \mathbb{R}^c$ for a generic per-token delta, i.e. a single row of $\mathbf{\Delta}_t$. All directional statistics in section 3 are computed at this per-token granularity.

3 ANALYSIS: THREE FAILURES OF RECURRENT STATE UPDATES

We systematically analyze the state update dynamics of recurrent 3D reconstruction models, our findings reveal three compounding failures, each deeper than the last, that together explain the progressive degradation observed in long-sequence inference.

3.1 OVER-UPDATE IS THE DOMINANT FAILURE MODE

We begin with a simple but revealing experiment beyond TTT3R. We scale every state update by a constant factor $\alpha < 1$, i.e. $\mathbf{S}_t = \mathbf{S}_{t-1} + \alpha\,\beta_t\,\mathbf{\Delta}_t$. This introduces no adaptivity, every update at every timestep is damped identically, yet it produces dramatic improvements (see Tab.1).

Table 1: Effect of constant dampening (α=0.5) on camera pose estimation (ATE↓). Improvements are consistent across datasets and sequence lengths, demonstrating that over-update is the dominant failure mode. Sintel sequences ($\sim$50 frames) are too short for over-update to accumulate.

	TUM		**ScanNet**		**Sintel**
Method	90f	1000f	90f	1000f	$\sim$50f
CUT3R	0.033	0.166	0.095	0.817	0.209
TTT3R	0.019	0.103	0.064	0.406	0.209
TTT3R + const. α=0.5	**0.016**	**0.079**	0.064	**0.283**	0.220
vs. CUT3R	−51%	−53%	−32%	−65%	+5%
vs. TTT3R	−15%	−23%	0%	−30%	+5%

A non-adaptive scalar improves every indoor setting except ScanNet at 90 frames, where TTT3R is already near-optimal, with reductions of 32–65% over CUT3R. The degradation it counteracts

compounds over time, with 1000f-vs-90f ATE ratios of $5.0\times$ on TUM and $8.6\times$ on ScanNet for CUT3R, so errors accumulate rather than self-correct. The exception is Sintel ($\sim$50 frames), where no method benefits, marking a lower bound on the length at which regulation matters. The issue is thus not when to dampen but that updates are always too large.

3.2 Adaptive Gates Collapse to Constants

TTT3R's confidence gate. TTT3R's per-token gate $\beta_t = \mathrm{sigmoid}(\sum_m \mathbf{Q}_{\mathbf{S}_{t-1}} \mathbf{K}_{\mathbf{X}_t}^\top)$ is intended to give low-confidence tokens smaller updates, but in practice exhibits almost no temporal variation. Across TUM, ScanNet, and the out-of-domain KITTI, the frame-averaged gate stays within $\mu \in [0.31, 0.35]$ at $\sigma_{\text{temporal}} \leq 0.03$ throughout (fig. 1b). Frame 1 is statistically indistinguishable from frame 1000. The collapse is structural. The summed cross-attention scores fall in the saturated regime of the sigmoid. Per-token spatial variation within each frame is, however, an order of magnitude larger ($\sigma_{\text{spatial}} \approx 0.15$), so β_t still encodes which tokens to dampen, just not when.

TTSA3R's TAUM gate. The concurrent TTSA3R Zheng et al. (2026) proposes an alternative gate, TAUM (Temporal Adaptive Update Modulation), built from the L2 norm of state changes: $g_t = \mathrm{sigmoid}\big(\|\Delta\mathbf{s}_t\|/\mathrm{mean}(\|\Delta\mathbf{s}_t\|) - \tau\big)$. Normalization by the running mean was meant to grant scale-invariant adaptivity, yet it is precisely the source of collapse. By construction $\|\Delta\mathbf{s}_t\|/\mathrm{mean}(\|\Delta\mathbf{s}_t\|)$ has mean 1.0, so the sigmoid input centers at $1.0 - \tau$ regardless of state dynamics. With the recommended $\tau = 1.5$, the gate locks to $\mathrm{sigmoid}(-0.5) = 0.378$ with $\sigma_{\text{temporal}} \approx 0.006$, a $3\times$ more severe collapse than TTT3R's, despite a completely different proximate cause.

This concerns the temporal gate only. TTSA3R fuses TAUM with a spatial module (SCUM) and is end-to-end a better gate than TTT3R's, improving on it in every setting we measure and ranking best overall on ScanNet at 90 frames (table 3). What the collapse predicts is not that TTSA3R is useless but that its advantage cannot survive accumulation: at 1000 frames it trails $\text{DDD3R}_{\text{ortho}}$ by 39.9% on TUM and $\text{DDD3R}_{\text{brake}}$ by 30.9% on ScanNet. A better scalar schedule delays the plateau without removing it.

Adaptivity does not predict improvement. If the residual fluctuation of the gate carried any timing signal, scenes with higher gate variance should benefit more. Across n=65 ScanNet scenes the correlation between $\mathrm{Var}(\beta_t)$ and per-scene ATE improvement is at most weakly *negative*, and unstable across split halves (section A.11). The signal, if any, points the wrong way, which rules out a useful timing role for the gate.

Why no scalar gate can help. The collapse of these particular gates is empirical, but a stronger statement holds for the whole family. Let $\boldsymbol{\delta}_t$ be the raw per-token delta and $\mathbf{d}_t$ any fixed reference direction. A scalar per-token gate applied as $\mathbf{S}_t = \mathbf{S}_{t-1} + g_t\,\boldsymbol{\delta}_t$ produces an applied update $\mathbf{u}_t = g_t\,\boldsymbol{\delta}_t$ that is collinear with $\boldsymbol{\delta}_t$, so for every $g_t \neq 0$

$$\cos^2(\mathbf{u}_t, \mathbf{d}_t) = \cos^2(\boldsymbol{\delta}_t, \mathbf{d}_t). \tag{5}$$

The drift energy of the applied update is therefore invariant to the gate, for any $\mathbf{d}_t$. A scalar gate rescales $\|\mathbf{u}_t\|$ only, so no schedule, however adaptive, alters the ratio of drift to novelty. Changing that ratio needs at least two coefficients, which is what eq. (6) introduces. This says what scalar gating cannot do, not that all scalar gates perform alike.

This explains why constant dampening (Failure 1) matches the nominally adaptive gates. Both are effectively constants, and the constants inherited from training are not the tuned optimum.

3.3 Directional Redundancy in State Updates

Failures 1 and 2 concern the magnitude of state updates. We now examine their content. Even with optimal magnitude control, a fundamental question remains. Does each update carry genuinely new information, or does it partially repeat the direction of prior updates?

Drift energy. To quantify directional redundancy, we track the cosine similarity between consecutive per-token deltas, $\cos(\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t-1})$, and define *drift energy* as its squared value, measuring the fraction of update energy aligned with the historical update direction.

Table 2: Directional statistics of state update deltas. Drift energy measures the fraction of update energy aligned with the historical direction. Both datasets exhibit substantial directional redundancy, but with different magnitudes reflecting different drift characteristics.

	TUM (8 seqs)	**ScanNet** (96 scenes)
$\cos(\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t-1})$	0.617 ± 0.037	0.767 ± 0.037
Drift energy ($\cos^2$)	0.398 ± 0.041	0.598 ± 0.054

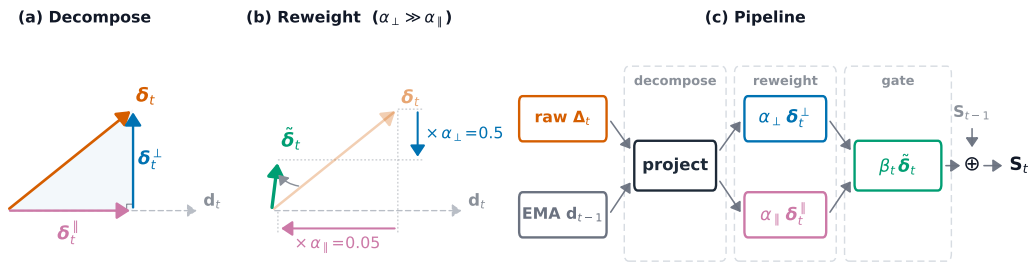

Figure 2: **DDD3R method overview.** **(a)** The raw per-token delta $\boldsymbol{\delta}_t$ is decomposed against the EMA-tracked drift direction $\mathbf{d}_t$ into a parallel component $\boldsymbol{\delta}_t^\parallel$ and an orthogonal complement $\boldsymbol{\delta}_t^\perp$. **(b)** What the two gains do to that same vector, at the default setting $\alpha_\perp = 0.5$, $\alpha_\parallel = 0.05$. The faded arrow repeats $\boldsymbol{\delta}_t$ and the dotted guides mark its extent across and along $\mathbf{d}_t$. The orthogonal extent is halved while the drift-aligned extent is cut to a twentieth, so the regulated delta $\tilde{\boldsymbol{\delta}}_t$ (green) leaves the drift direction almost untouched and points nearly perpendicular to it. **(c)** The pipeline. The raw delta $\boldsymbol{\Delta}_t$ and the previous EMA drift $\mathbf{d}_{t-1}$ enter the decompose stage, the two components are reweighted, and the result is gated by the base-model token mask β_t before being added to $\mathbf{S}_{t-1}$.

Table 2 reports these statistics across datasets. In both datasets, a large fraction of each update is directionally redundant, aligned with the direction the state has already been moving. This structure is invisible to scalar gates, which attenuate all directions equally.

The dual nature of drift. A critical finding is that drift energy alone does not determine whether suppression helps. On TUM ($\bar{e}\approx0.40$) aggressive suppression of the drift component is best, so drift there mostly reflects harmful error accumulation. On ScanNet ($\bar{e}\approx0.60$) the same suppression degrades performance, because the higher drift energy reflects useful geometric refinement. We verify this on 90 valid ScanNet scenes (fig. 7b, section A.11). The correlation between drift energy and DDD3R's ATE improvement is weak but consistently negative ($r=-0.25$, $p=0.018$), with same-sign split halves and a bootstrap 95% CI of $[-0.40, -0.09]$ excluding zero. We do not claim a strong predictor. Drift energy gives a consistent signed signal of when directional decomposition hurts, whereas constant dampening shows no significant correlation with it, since scalar methods cannot exploit this structure at all.

4 Method

Failures **M1**–**M3** in section 3 jointly demand a regulator that is both directional, since scalar gates cannot exploit drift redundancy, and adaptive to per-scene drift characteristics. We present **DDD3R** (**D**irectional **D**ecomposition and **D**ampening for Recurrent **3**D **R**econstruction), a training-free plug-in that satisfies both with zero parameters and zero added memory. Adopting the notation of section 2, DDD3R replaces the inner delta $\boldsymbol{\Delta}_t$ in eq. (4) with a regulated version while retaining the base-model gate β_t as the outer per-token mask.

4.1 DDD3R: A UNIFIED UPDATE RULE

DDD3R modifies the state update in three stages, decompose, reweight and gate, each directly motivated by a diagnostic finding.

$$\boxed{\mathbf{S}_t = \mathbf{S}_{t-1} + \beta_t \left(\alpha_\perp \, \boldsymbol{\delta}_t^\perp + \alpha_\parallel \, \boldsymbol{\delta}_t^\parallel \right)} \tag{6}$$

where $\boldsymbol{\delta}_t^\parallel$ and $\boldsymbol{\delta}_t^\perp$ are the drift and orthogonal components of the raw per-token delta (section 4.1.1), $\alpha_\parallel$ and $\alpha_\perp$ are scalar reweighting coefficients (section 4.1.2), and β_t is the original per-token gate inherited from the base model (eq. (4)).

Equation (6) subsumes all previously discussed approaches as special cases of a single parameterized family (section 4.2). In the remainder of this section, we define each component and explain how it addresses the corresponding diagnostic finding.

4.1.1 STAGE 1: DIRECTIONAL DECOMPOSITION (ADDRESSING M3)

To exploit the scene-dependent directional overlap quantified in section 3.3, we decompose each per-token delta into a drift-aligned component and an orthogonal complement. We track the drift direction via an exponential moving average over unit-normalized deltas:

$$\mathbf{d}_t = \beta_{ema} \, \mathbf{d}_{t-1} + (1 - \beta_{ema}) \, \hat{\boldsymbol{\delta}}_t \tag{7}$$

where $\hat{\boldsymbol{\delta}}_t = \boldsymbol{\delta}_t / \|\boldsymbol{\delta}_t\|$ is the unit-normalized delta and β_{ema} controls the tracking momentum. The drift direction $\mathbf{d}_t$ captures the smoothed historical trajectory of the state in feature space.

The raw delta is then decomposed via projection onto $\mathbf{d}_t$:

$$\boldsymbol{\delta}_t^\parallel = \frac{\langle \boldsymbol{\delta}_t, \, \mathbf{d}_t \rangle}{\|\mathbf{d}_t\|^2} \, \mathbf{d}_t \qquad \text{(projection onto drift direction)} \tag{8}$$

$$\boldsymbol{\delta}_t^\perp = \boldsymbol{\delta}_t - \boldsymbol{\delta}_t^\parallel \qquad \text{(orthogonal complement)} \tag{9}$$

The drift component $\boldsymbol{\delta}_t^\parallel$ lies along the direction the state has been moving. The orthogonal component $\boldsymbol{\delta}_t^\perp$ is perpendicular to this history and carries information not redundant with prior updates.

4.1.2 STAGE 2: DIFFERENTIAL REWEIGHTING (ADDRESSING M1)

DDD3R applies separate coefficients to the two components to address both M1 and M3 simultaneously:

$$\tilde{\boldsymbol{\delta}}_t = \alpha_\perp \, \boldsymbol{\delta}_t^\perp + \alpha_\parallel \, \boldsymbol{\delta}_t^\parallel, \qquad 0 < \alpha_\parallel \leq \alpha_\perp \leq 1. \tag{10}$$

The constraint $\alpha_\parallel \leq \alpha_\perp$ encodes the design principle that redundant drift content is at least as expendable as novel orthogonal content. We default to $\alpha_\perp$=0.5, $\alpha_\parallel$=0.05 from a light TUM grid search, and sensitivity is reported in section 5.2.

4.1.3 STAGE 3: PER-TOKEN GATING (INHERITED, ADDRESSING M2)

The regulated delta $\tilde{\boldsymbol{\delta}}_t$ is multiplied by the base model's per-token gate β_t before application:

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \beta_t \, \tilde{\boldsymbol{\delta}}_t. \tag{11}$$

Although our analysis (**M2**) shows that β_t collapses to near-constant values, we retain it rather than discard it. It still provides meaningful per-token spatial selectivity ($\sigma_{\text{spatial}} \gg \sigma_{\text{temporal}}$), and removing it would require retraining. DDD3R operates within the update that β_t already modulates, adding directional awareness without interfering with the spatial structure the base model has learned.

This separates two orthogonal dimensions of state update control (fig. 2): $\beta_t \in \mathbb{R}^{n \times 1}$ governs token-level update intensity, while $(\alpha_\perp, \alpha_\parallel)$ govern direction-level composition within each token's update. A head-to-head comparison with prior update rules is given in table 9 (section A.7).

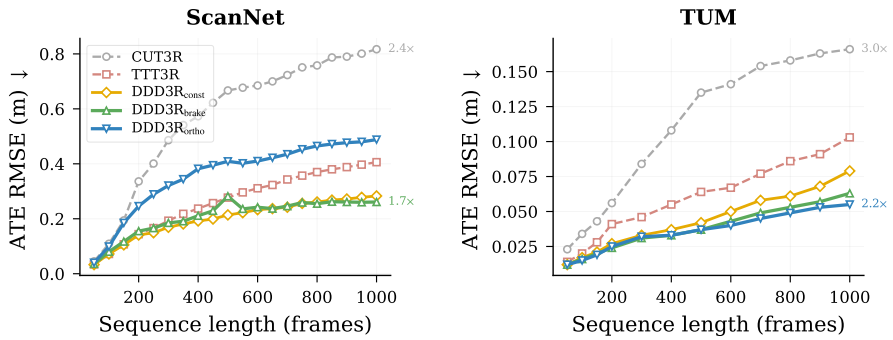

Figure 3: **Relative pose error (ATE) as a function of sequence length on ScanNet and TUM.** All methods are comparable at short sequences ($\leq$100 frames) but diverge dramatically as length grows, confirming that over-update is a cumulative phenomenon. Growth ratios (200f$\rightarrow$1000f) shown at right. On ScanNet, DDD3R$_{\text{brake}}$ achieves the lowest growth (1.7$\times$), and on TUM, DDD3R$_{\text{ortho}}$ achieves 2.2$\times$ vs. CUT3R's 3.0$\times$. The dataset-dependent ranking reversal is a central empirical finding (section A.4).

4.2 A Continuous Spectrum of Instantiations

Equation (6) subsumes prior update strategies as boundary cases on a single axis, how much to differentiate drift from novelty, governed by the gap $\alpha_\perp - \alpha_\parallel$. **Constant dampening** ($\alpha_\perp = \alpha_\parallel = \alpha$) collapses the decomposition to $\alpha\boldsymbol{\delta}_t$, recovering isotropic magnitude control (M1) without directional awareness, and the strong constant baseline in our experiments is exactly this special case. **Full directional decomposition** ($\alpha_\perp \gg \alpha_\parallel$) is the default DDD3R operating point and additionally addresses M3 by suppressing drift more aggressively than novelty, which matters most on moderate-drift scenes such as TUM. A **temporal brake** $g_t = \text{sigmoid}(-\tau \cos(\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t-1}))$ recovers the same qualitative behaviour with a single scalar per token. When consecutive deltas align (cos high), the delta is predominantly drift and the brake attenuates the whole update, and when they disagree, the brake passes the update through. Section A.8 shows that brake is the scalar relaxation of eq. (6). It tracks DDD3R's effective magnitude $\sqrt{\alpha_\perp^2 - (\alpha_\perp^2 - \alpha_\parallel^2)\cos^2\theta}$ qualitatively in the empirical regime $\cos\theta \in [0.4, 0.8]$, but cannot apply differentiated suppression per direction because it only has one knob. DDD3R refines this by (i) tracking the drift direction with EMA over the full history rather than the last frame, and (ii) assigning separate coefficients to drift and orthogonal components. A **drift-energy-adaptive** extension ($\gamma > 0$) interpolates $\alpha_\parallel$ per token along the spectrum. We treat it as a spectrum-exploration tool rather than a recommended configuration and defer details to sections A.3 and A.4.

5 Experiments

We evaluate three operating points of the DDD3R framework, namely DDD3R$_{\text{const}}$ ($\alpha_\perp = \alpha_\parallel$), DDD3R$_{\text{brake}}$ (the scalar relaxation, see section A.8) and DDD3R$_{\text{ortho}}$ ($\alpha_\perp \gg \alpha_\parallel$, default), against CUT3R Wang et al. (2025b) and TTT3R Chen et al. (2026), all sharing the same pretrained checkpoint with no retraining. We evaluate relative pose on TUM Sturm et al. (2012), ScanNet Dai et al. (2017), Sintel Butler et al. (2012) and KITTI Geiger et al. (2012), video depth on KITTI, Bonn Palazzolo et al. (2019) and Sintel, and 3D reconstruction on 7Scenes Shotton et al. (2013). Metrics are ATE RMSE after Sim(3) for pose, scale-and-shift-invariant abs_rel for depth, Acc/Comp/NC for reconstruction. Full configurations and per-dataset protocols in section A.2.

5.1 Main Results

Relative pose estimation. Figure 3 shows the full scaling behavior, and table 3 reports representative operating points. On **long sequences (1000 frames)**, all DDD3R variants substantially improve over CUT3R and TTT3R. On TUM, DDD3R$_{\text{ortho}}$ achieves the best result (ATE 0.055, -67% vs. CUT3R), while on ScanNet DDD3R$_{\text{brake}}$ performs best (0.261, -68%). This divergence

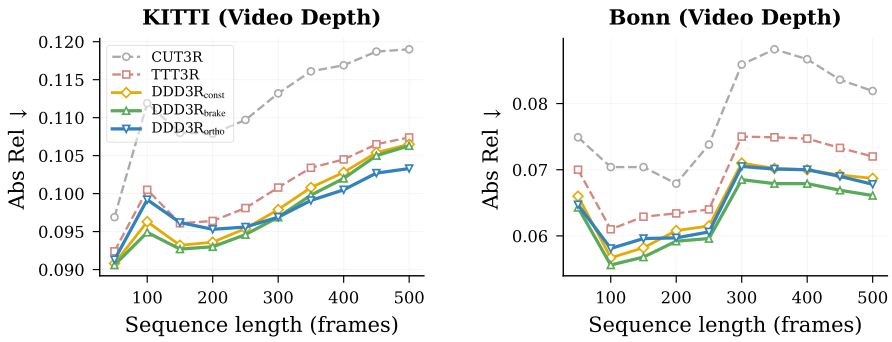

Figure 4: **Video depth (abs_rel ↓) vs. sequence length.** On KITTI (left), DDD3R$_{brake}$ leads at short lengths and DDD3R$_{ortho}$ takes over beyond ∼300 frames. On Bonn (right), DDD3R$_{brake}$ leads consistently. Both crossovers are explained by drift-energy differences (section 3.3).

Table 3: **Relative pose ATE** (m, ↓).

	TUM		ScanNet		Sintel	KITTI
Method	90f	1000f	90f	1000f	∼50f	full
CUT3R Wang et al. (2025b)	0.033	0.166	0.095	0.817	**0.209**	192.84
TTT3R Chen et al. (2026)	0.019	0.103	0.064	0.406	**0.209**	161.14
TTSA3R Zheng et al. (2026)	0.016	0.091	**0.058**	0.378	**0.209**	—
DDD3R$_{const}$	0.016	0.079	0.064	0.283	0.220	156.61
DDD3R$_{brake}$	0.015	0.063	0.072	**0.261**	0.237	166.89
DDD3R$_{ortho}$	**0.015**	**0.055**	0.087	0.488	0.236	**154.10**

Table 4: **Video depth abs_rel** (↓).

Method	**KITTI**	**Bonn**	**Sintel**
CUT3R	0.119	0.082	0.465
TTT3R	0.107	0.072	0.433
TTSA3R	0.106	0.072	0.425
DDD3R$_{const}$	0.107	0.069	0.408
DDD3R$_{brake}$	0.106	**0.066**	**0.402**
DDD3R$_{ortho}$	**0.103**	0.068	0.418

is consistent with the drift-energy diagnosis (section 3.3). TUM's moderate drift ($\bar{e}$≈0.40) benefits from directional decomposition, while ScanNet's high drift ($\bar{e}$≈0.60) is better served by isotropic dampening. On **short sequences (90 frames)**, the gain is dataset-dependent. TUM already shows −55% with DDD3R$_{ortho}$ (0.015 vs. 0.033), but on ScanNet 90f no DDD3R variant beats TTT3R, since over-update has not yet accumulated enough at 90 frames in higher-drift scenes (section 3.1). On **Sintel (∼50 frames)**, no dampening method improves over CUT3R/TTT3R. This is expected, since the model is trained on short pairs and per-step magnitudes are already well-calibrated at this length.

Video depth prediction. Figure 4 reports abs_rel as a function of sequence length on KITTI and Bonn, and table 4 summarizes representative operating points. Every DDD3R variant strictly improves over both baselines on all three benchmarks. On KITTI a crossover emerges near ∼300 frames, with DDD3R$_{brake}$ leading at short lengths and DDD3R$_{ortho}$ beyond, while on Bonn DDD3R$_{brake}$ leads consistently.

3D reconstruction and cross-dataset trade-off. On 7Scenes, DDD3R$_{brake}$ reaches Acc 0.021 (−77% vs. CUT3R) and DDD3R$_{ortho}$ Acc 0.026, Comp 0.022 (full table in section A.6). Across all four tasks no single regulated configuration dominates everywhere. DDD3R$_{brake}$ leads on ScanNet (relpose) and Bonn (vdepth) and DDD3R$_{ortho}$ on TUM and KITTI, a ranking reversal that section A.4 traces to a continuous γ sweep (section A.3) where drift energy, a scene-intrinsic property, governs the appropriate operating point.

5.2 Ablation Studies and Practical Recommendations

Hyperparameter sensitivity. On TUM, $\alpha_{\parallel} \in [0.05, 0.20]$ varies ATE by <2% and β_{ema}=0.95 is best; on ScanNet the optimum drifts toward larger $\alpha_{\parallel}$, consistent with its higher drift energy.

Practical recommendations. The experiments yield a clear deployment guideline. **(i) DDD3R$_{brake}$ is the most robust default**, ranking 1st on ScanNet 1000f (relpose), Bonn and Sintel (vdepth), and 7Scenes (recon) without any directional decomposition. **(ii) DDD3R$_{ortho}$ gives an additional**

10–15% on low-drift indoor scenes (TUM and similar slow-scanning settings), where drift mostly reflects harmful error accumulation. **(iii) Default DDD3R$_{\text{ortho}}$ (γ=0) is the best aggregate for outdoor driving.** KITTI Odometry is also our strongest test of domain shift, since the checkpoint is trained on short indoor-heavy pairs while these are outdoor driving sequences of 271–4661 frames. DDD3R$_{\text{ortho}}$ reduces ATE by 20.1% over CUT3R across the 11 sequences, and on the official segment-based metrics it reduces translation error by 7.6% and rotation error by 58.6%, and per-sequence optima drift toward mild $\gamma \in [0.5, 2]$ (table 5). Switching between operating points changes only scalar hyperparameters.

Beyond this backbone. Two further studies are in the appendix. On **ARKitScenes** (section A.10), 18 real-world handheld scenes, drift energy measured before any pose evaluation correctly predicted which operating point would fail. On **Point3R** Wu et al. (2026), a different backbone with an explicit pointer memory (section A.9), the magnitude fix gives 5.4% on ScanNet at the same α=0.5, while its low drift energy correctly indicates that directional decomposition should not be used there.

Computational overhead. DDD3R adds zero parameters and zero peak memory and runs at 9.95 FPS versus CUT3R's 8.44 FPS on TUM. The full breakdown is in section A.6.

6 Related Work

Feed-forward and recurrent 3D reconstruction. Feed-forward methods regress 3D geometry from image pairs or small sets Wang et al. (2024); Leroy et al. (2024); Wang et al. (2025a), building on classical SfM/MVS Schönberger & Frahm (2016); Schönberger et al. (2016); Yao et al. (2018) and learned depth/scene representations Yang et al. (2024); Piccinelli et al. (2024); Mildenhall et al. (2020); Kerbl et al. (2023), but cannot accumulate state. CUT3R Wang et al. (2025b) extends this to streaming video via a persistent cross-attention state, TTT3R Chen et al. (2026) adds an attention-derived per-token learning rate, and TTSA3R Zheng et al. (2026) adds temporal (TAUM) and spatial (SCUM) gates. All three are calibrated on short training pairs and degrade recurrently, the failure we diagnose in section 3.

Reducing redundancy in the state. A parallel line attacks redundancy along the spatial-token axis, that is, how many tokens carry the state. ZPressor Wang et al. (2025d) compresses it through an information bottleneck, Long-LRM Ziwen et al. (2025) merges tokens, and iLRM Kang et al. (2026) refines a decoupled representation iteratively, with a recent survey Wang et al. (2026) organising the space. These methods compress the width of the state. DDD3R instead regulates its trajectory. The two axes are composable, since a compressed state still accumulates drift and eq. (6) would apply to it unchanged.

Test-time training, gating, and update regulation. Test-time training Sun et al. (2020); Liu et al. (2021); Sun et al. (2025); Wang et al. (2025c) confronts stability–plasticity Parisi et al. (2019). Prior remedies modulate how much to update via learning-rate schedules Wang et al. (2021), confidence thresholds Wang et al. (2022), entropy minimization Liang et al. (2020), or scalar gates Hochreiter & Schmidhuber (1997); Cho et al. (2014). Closest in spirit are continual-learning gradient projections orthogonal to past knowledge Kirkpatrick et al. (2017); Lopez-Paz & Ranzato (2017); Farajtabar et al. (2020), but those act on parameter gradients during training, not state deltas at inference. DDD3R brings directional decomposition to inference-time recurrent state regulation, replacing scalar magnitude control with explicit drift–novelty separation.

7 Conclusion

We presented **DDD3R**, a training-free regulator for the state update of recurrent 3D reconstruction. We diagnosed three compounding failures, uncontrolled magnitude, gate collapse and drift redundancy, and showed that no scalar gate can reach the last of them. A single update rule that decomposes each per-token delta and reweights the components differentially addresses all three, with constant dampening and temporal braking as boundary cases. Drift energy, a scene-intrinsic property, governs which operating point is appropriate. **Limitations.** Per-frame adaptive predictors saturate at the fixed-γ Pareto frontier, since the right operating point is a sequence-level property; selecting it online remains open.

AI USE STATEMENT

We used a generative AI assistant (Anthropic's Claude, accessed through a coding agent) during this project. Its use is disclosed below.

Tasks with required disclosure, used. We used it to design and give feedback on research methodology and experiments, including the paired significance tests reported throughout section 5. We used it to refine hypotheses, sharpening the drift-energy hypothesis of section 3 into the quantitative predictions tested in section A.10. We used it to interpret results, in reading the per-scene and scaling measurements.

Tasks with required disclosure, not used. We did not use generative AI to generate synthetic data sets, to clean or reformat datasets, to develop theoretical models or conceptual frameworks beyond the single claim named above, or to assist with translation. Formulating questions for surveys or interviews and transcribing recordings of research material are not applicable to this work.

Tasks with recommended disclosure. We also used generative AI to draft and edit parts of this paper, to create its figures, to identify and summarise related literature, to format references, and to search for information.

Review procedures. We have reviewed all AI-assisted work. All AI-written code was read by the authors, and every reported number was regenerated from raw evaluation output rather than transcribed. The proof of eq. (5) was re-derived independently by the authors. Every bibliographic entry in this paper was verified against a public authoritative record, namely the publisher's page, the official proceedings, or arXiv, checking the title, the complete author list, the venue and the year. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies inference-time regulation of the state update in existing, publicly released 3D reconstruction models, and introduces no new data collection. All datasets used (TUM RGB-D, ScanNet, Sintel, KITTI, Bonn RGB-D Dynamic, 7-Scenes, ARKitScenes) are established public benchmarks used under their respective licences. Some contain recordings of people; we use them only for geometric evaluation and perform no person identification. We are not aware of ethical concerns specific to this work beyond those already inherent to the underlying benchmarks.

REPRODUCIBILITY STATEMENT

DDD3R is training-free and adds no parameters, so reproduction requires only the publicly released base checkpoint. Section 4 gives the complete update rule, Appendix A.1 gives the default hyperparameters and the algorithm in pseudocode, and Appendix A.2 gives the per-dataset evaluation protocols, sequence lists and metric definitions. All reported numbers come from a single evaluation script over the released checkpoints. We will release the code and the evaluation configurations.

REFERENCES

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, 2012.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. In *International Conference on Learning Representations*, volume 2026, pp. 50694–50718, 2026.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Gyeongjin Kang, Seungtae Nam, Seungkwon Yang, Xiangyu Sun, Sameh Khamis, Abdelrahman Mohamed, and Eunbyung Park. ilrm: An iterative large 3d reconstruction model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37332–37342, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, 2020.

Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. TTT++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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. In *European Conference on Computer Vision (ECCV)*, 2020.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10106–10116. IEEE, 2024.

Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. In *International Conference on Machine Learning (ICML)*, 2025.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Renhao Wang, Yu Sun, Arnuv Tandon, Yossi Gandelsman, Xinlei Chen, Alexei A Efros, and Xiaolong Wang. Test-time training on video streams. *Journal of Machine Learning Research*, 26(9):1–29, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Weijie Wang, Donny Y. Chen, Zeyu Zhang, Duochao Shi, Akide Liu, and Bohan Zhuang. ZPressor: Bottleneck-aware compression for scalable feed-forward 3DGS. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025d.

Weijie Wang, Qihang Cao, Sensen Gao, Donny Y. Chen, Haofei Xu, Wenjing Bian, Songyou Peng, Tat-Jen Cham, Chuanxia Zheng, Andreas Geiger, Jianfei Cai, Jia-Wang Bian, and Bohan Zhuang. Feed-forward 3D scene modeling: A problem-driven perspective. *arXiv preprint arXiv:2604.14025*, 2026.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *Advances in Neural Information Processing Systems*, 38:69675–69699, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381. IEEE, 2024.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European conference on computer vision*, pp. 785–801. Springer, 2018.

Zhijie Zheng, Xinhao Xiang, and Jiawei Zhang. TTSA3R: Training-free temporal-spatial adaptive persistent state for streaming 3D reconstruction. *arXiv preprint arXiv:2601.22615*, 2026.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, volume 2026, pp. 88055–88072, 2026.

Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-LRM: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

A Appendix

A.1 Implementation Details

DDD3R is a drop-in modification to the state update step of any recurrent 3D reconstruction model that emits a per-token candidate state. Algorithm 1 summarizes the complete procedure used in all experiments.

Algorithm 1 DDD3R State Update

Require: State $\mathbf{S}_{t-1} \in \mathbb{R}^{n\times c}$, candidate $\hat{\mathbf{S}}_t$, per-token gate $\beta_t \in \mathbb{R}^{n\times 1}$
Require: EMA drift direction $\mathbf{d}_{t-1} \in \mathbb{R}^{n\times c}$ (initialized to $\mathbf{0}$)
Require: Hyperparameters $\alpha_\perp$, $\alpha_\parallel$, β_{ema}, γ

$\boldsymbol{\delta}_t \leftarrow \hat{\mathbf{S}}_t - \mathbf{S}_{t-1}$ ▷ Raw delta
$\hat{\boldsymbol{\delta}}_t \leftarrow \boldsymbol{\delta}_t \,/\, \|\boldsymbol{\delta}_t\|$ ▷ Unit-normalized direction
$\mathbf{d}_t \leftarrow \beta_{ema}\, \mathbf{d}_{t-1} + (1-\beta_{ema})\, \hat{\boldsymbol{\delta}}_t$ ▷ Update EMA direction
$\boldsymbol{\delta}_t^{\parallel} \leftarrow \frac{\langle \boldsymbol{\delta}_t,\, \mathbf{d}_t\rangle}{\|\mathbf{d}_t\|^2}\, \mathbf{d}_t$ ▷ Drift component
$\boldsymbol{\delta}_t^{\perp} \leftarrow \boldsymbol{\delta}_t - \boldsymbol{\delta}_t^{\parallel}$ ▷ Orthogonal component
if $\gamma > 0$ **then** ▷ Drift-energy-adaptive mode
    $e_t \leftarrow \langle \hat{\boldsymbol{\delta}}_t,\, \hat{\mathbf{d}}_t\rangle^2$ where $\hat{\mathbf{d}}_t = \mathbf{d}_t / \|\mathbf{d}_t\|$
    $\alpha_{\parallel}^{(t)} \leftarrow e_t^{\gamma}\, \alpha_\perp + (1 - e_t^{\gamma})\, \alpha_\parallel$
else
    $\alpha_{\parallel}^{(t)} \leftarrow \alpha_\parallel$
end if
$\tilde{\boldsymbol{\delta}}_t \leftarrow \alpha_\perp\, \boldsymbol{\delta}_t^{\perp} + \alpha_{\parallel}^{(t)}\, \boldsymbol{\delta}_t^{\parallel}$ ▷ Reweighted delta
$\mathbf{S}_t \leftarrow \mathbf{S}_{t-1} + \beta_t\, \tilde{\boldsymbol{\delta}}_t$ ▷ Gated update
return $\mathbf{S}_t$, $\mathbf{d}_t$

Computational overhead. The module stores only the EMA direction $\mathbf{d}_t \in \mathbb{R}^{n\times c}$, which has the same shape as a single state and is maintained in-place. All operations are element-wise or per-token inner products, adding negligible computation. Empirically, DDD3R runs at 9.95 FPS versus CUT3R's 8.44 FPS on a single Nvidia A6000 GPU (section 5.2); the dampened update reduces memory write pressure.

Hyperparameters. DDD3R has four hyperparameters: $\alpha_\perp$, $\alpha_\parallel$, β_{ema}, γ. We use $(\alpha_\perp, \alpha_\parallel, \beta_{ema}, \gamma) = (0.5, 0.05, 0.95, 0)$ as the default configuration throughout all experiments unless otherwise noted. Sensitivity analysis (section 5.2) shows $\alpha_\parallel \in [0.05, 0.2]$ varies ATE by $<2\%$

on TUM, $\alpha_\perp \geq 0.5$ is stable, and $\beta_{ema} = 0.95$ is consistently optimal. The drift-adaptive mode ($\gamma > 0$) is explored as an ablation; the fixed operating point (γ=0) already achieves strong results across all evaluation tasks.

A.2 Experimental Setup

Baselines and DDD3R variants. We compare two baselines and three operating points along the DDD3R spectrum (section 4.2), all sharing the same pretrained base model with no retraining: (i) **CUT3R** Wang et al. (2025b): full state replacement ($\beta_t \equiv 1$). (ii) **TTT3R** Chen et al. (2026): attention-derived per-token gate. (iii) DDD3R$_{\text{const}}$ ($\alpha_\perp$=$\alpha_\parallel$=0.5, γ=0): isotropic dampening, the $\alpha_\perp$=$\alpha_\parallel$ degenerate case of eq. (6). (iv) DDD3R$_{\text{brake}}$ (τ=2): the temporal-brake scalar relaxation of eq. (6) (see section A.8). (v) DDD3R$_{\text{ortho}}$ ($\alpha_\perp$=0.5, $\alpha_\parallel$=0.05, β_{ema}=0.95, γ=0): default full directional decomposition.

Datasets and protocols. For *relative pose estimation*, we evaluate on TUM Sturm et al. (2012) (8 sequences), ScanNet Dai et al. (2017) (96 scenes, 65–90 valid after filtering), Sintel Butler et al. (2012) (14 sequences, ∼50 frames), and KITTI Odometry Geiger et al. (2012) (11 sequences, 271–4661 frames, outdoor driving). Following prior work Wang et al. (2025b); Chen et al. (2026), we report ATE RMSE (meters) after Sim(3) alignment. Indoor sequences are evaluated at both 90 and 1000 frames to assess length dependence; KITTI uses full-length sequences. For *video depth prediction*, we evaluate on KITTI Geiger et al. (2012), Bonn Palazzolo et al. (2019), and Sintel Butler et al. (2012), reporting scale-and-shift-invariant absolute relative error (abs_rel ↓). For *3D reconstruction*, we evaluate on 7Scenes Shotton et al. (2013), reporting accuracy (Acc ↓), completeness (Comp ↓), and normal consistency (NC ↑).

A.3 Drift-Energy-Adaptive Interpolation ($\gamma > 0$)

The default DDD3R operating point uses γ=0, which fixes $\alpha_\parallel$ at its tuned value. We also explored a drift-energy-adaptive variant that interpolates per-token between full directional decomposition and isotropic dampening based on the instantaneous alignment between the current delta and the EMA drift:

$$e_t = \langle \hat{\boldsymbol{\delta}}_t, \hat{\mathbf{d}}_t \rangle^2, \qquad \alpha_\parallel^{(t)} = e_t^\gamma \, \alpha_\perp + (1 - e_t^\gamma) \, \alpha_\parallel \tag{12}$$

where $\hat{\mathbf{d}}_t = \mathbf{d}_t / \|\mathbf{d}_t\|$. When drift energy e_t is high, $e_t^\gamma \to 1$ and $\alpha_\parallel^{(t)} \to \alpha_\perp$, collapsing the framework toward isotropic dampening; when low, $e_t^\gamma \to 0$ and full directional selectivity is retained. Setting γ=0 recovers the fixed operating point. We treat γ>0 as an ablation tool that exposes the continuous spectrum rather than a recommended deployment configuration.

A.4 DDD3R Spectrum and Cross-Dataset Ranking Reversal

A claim made informally in section 5.1 is that all reasonable update strategies sit on a continuous spectrum controlled by the gap $\alpha_\perp - \alpha_\parallel$ and the drift-adaptive parameter γ (section A.3). Table 5 validates this by sweeping γ from 0 toward larger values that progressively interpolate toward isotropic dampening.

Table 5: **The DDD3R spectrum.** Sweeping γ traces a continuous path from directional decomposition (γ=0) toward isotropic dampening ($\gamma \to \infty$). Different datasets favor different operating points, consistent with their drift-energy profiles. Relative pose ATE ↓.

DDD3R config	TUM 1000f	ScanNet 1000f	KITTI 04	KITTI 03
DDD3R$_{\text{const}}$	0.079	<u>0.283</u>	<u>5.10</u>	60.68
DDD3R$_{\text{brake}}$	0.063	**0.261**	7.05	<u>25.38</u>
DDD3R$_{\text{ortho}}$ ($\gamma = 0$)	0.055	0.488	17.79	39.91
DDD3R$_{\text{ortho}}$ ($\gamma = 0.5$)	0.061	0.327	**4.89**	46.54
DDD3R$_{\text{ortho}}$ ($\gamma = 1$)	0.056	0.358	6.92	37.89
DDD3R$_{\text{ortho}}$ ($\gamma = 2$)	<u>0.055</u>	0.394	8.78	**24.58**
DDD3R$_{\text{ortho}}$ ($\gamma = 3$)	**0.054**	0.407	8.14	26.79

The spectrum reveals a systematic pattern that aligns with the diagnostic findings. **TUM** ($\bar{e}\approx 0.40$, moderate drift) favors the directional end: $\gamma \geq 2$ achieves ATE ≤ 0.055, with $\gamma{=}3$ reaching the overall best (0.054). **ScanNet** ($\bar{e}\approx 0.60$, high drift) favors the isotropic end: DDD3R$_{\text{const}}$ (0.283) and DDD3R$_{\text{brake}}$ (0.261) outperform every directional variant. **KITTI** sequences split: seq04 favors mild decomposition ($\gamma{=}0.5$), while seq03 favors strong drift adaptation ($\gamma{=}2$, with DDD3R$_{\text{brake}}$ a close second). No single γ dominates across all datasets.

The cross-dataset ranking systematically reverses (table 6): DDD3R$_{\text{brake}}$ ranks 1st on ScanNet but 4th on KITTI, while DDD3R$_{\text{ortho}}$ ($\gamma{=}0$) ranks 1st on TUM and KITTI but 4th on ScanNet. We read this as supporting evidence for the central empirical claim that drift energy—a scene-intrinsic property—governs which operating point along the spectrum is appropriate.

Table 6: **Cross-dataset method ranking.** No fixed configuration is universally optimal; rankings reverse systematically with drift energy. KITTI ranks aggregate ATE over the 11 KITTI Odometry sequences (table 3); the per-sequence picture is more nuanced (table 5).

Method	TUM	ScanNet	KITTI	Drift regime
DDD3R$_{\text{const}}$	3rd	2nd	2nd	Direction-agnostic
DDD3R$_{\text{brake}}$	2nd	**1st**	4th	High-drift optimal
DDD3R$_{\text{ortho}}$ ($\gamma{=}0$)	**1st**	4th	**1st**	Low-drift optimal
DDD3R$_{\text{ortho}}$ ($\gamma \in [0.5, 2]$)	1st	3rd	per-seq best	Mild adaptation

A.5 Qualitative Results

We provide two qualitative comparisons that illustrate the long-sequence failure mode and how DDD3R corrects it. Figure 5 overlays the camera trajectories produced by CUT3R, TTT3R, and DDD3R$_{\text{ortho}}$ on a 1000-frame ScanNet sequence: CUT3R drifts off the reference path early and never recovers, TTT3R reduces but does not eliminate the drift, and DDD3R$_{\text{ortho}}$ stays close to ground truth throughout. Figure 6 shows per-frame depth predictions on a Bonn sequence: CUT3R's predictions degrade as the recurrent state accumulates updates over time (later frames noticeably noisier), while DDD3R$_{\text{ortho}}$ maintains stable, geometrically consistent depth across the full sequence. Both visualizations confirm that the gains reported in tables 3 and 4 are not driven by averaging effects but reflect a per-frame qualitative improvement.

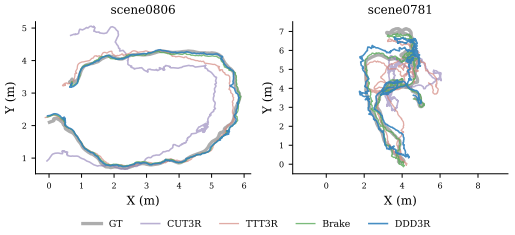

Figure 5: **Trajectory comparison on ScanNet (1000f).** CUT3R drifts severely; DDD3R closely tracks the reference path.

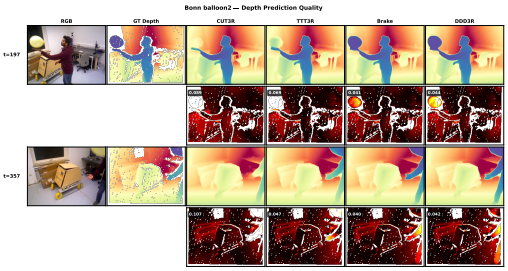

Figure 6: **Depth prediction on Bonn.** CUT3R degrades over time; DDD3R maintains stable depth.

A.6 Inference Overhead and 3D Reconstruction

Inference overhead. Table 7 reports inference speed and peak memory on TUM (3 sequences $\times$ 200 frames, averaged over 3 runs). DDD3R adds zero parameters and zero peak memory; all regulated methods run slightly faster than CUT3R, likely because dampened updates reduce memory write pressure.

3D reconstruction (7Scenes). All regulated methods substantially improve accuracy and completeness over CUT3R; differences among variants are smaller than on the relpose / vdepth benchmarks,

Table 7: **Inference overhead.** DDD3R adds zero parameters and zero peak memory; throughput is comparable to or better than CUT3R.

Method	FPS ↑	Peak Mem (GB)	vs. CUT3R
CUT3R	8.44	6.14	—
TTT3R	9.82	6.14	+16%
$\text{DDD3R}_{\text{brake}}$	10.03	6.14	+19%
$\text{DDD3R}_{\text{ortho}}$	9.95	6.14	+18%

reflecting the shorter effective sequence lengths in 7Scenes (∼500–1000 frames with overlapping views), where magnitude control alone captures most of the benefit.

Table 8: **3D reconstruction on 7Scenes.** All regulated methods substantially improve accuracy and completeness over CUT3R.

Method	Acc ↓	Comp ↓	NC ↑
CUT3R	0.092	0.048	0.563
TTT3R	0.027	0.023	**0.581**
$\text{DDD3R}_{\text{brake}}$	**0.021**	**0.022**	<u>0.580</u>
$\text{DDD3R}_{\text{ortho}}$	<u>0.026</u>	**0.022**	0.577

A.7 Comparison of State Update Rules

Table 9 places DDD3R alongside the three update rules it generalizes. CUT3R is the unregulated baseline, applying the full cross-attention delta at every step ($\beta_t \equiv 1$). TTT3R Chen et al. (2026) introduces a per-token sigmoid gate driven by cross-attention confidence, and TTSA3R Zheng et al. (2026) composes a temporal (TAUM) and a spatial (SCUM) gate from state-change norms. All three operate purely on update *magnitude*: they multiply the delta by a scalar (per-token, but scalar in the channel dimension) and leave the directional structure of $\mathbf{\Delta}_t$ untouched. DDD3R adds a second axis of control by replacing the delta itself with a directionally reweighted version, $\alpha_\perp \boldsymbol{\delta}_t^\perp + \alpha_\parallel \boldsymbol{\delta}_t^\parallel$, while inheriting the base-model gate β_t unchanged. This is what allows it to address **M3** (drift redundancy) on top of **M1** (magnitude); the prior rules are recovered as boundary cases of eq. (6) and add no learnable parameters either.

Table 9: Comparison of state update rules. DDD3R generalizes prior approaches by introducing directional control orthogonal to the scalar learning rate. All methods share the same base architecture and add no learnable parameters.

Method	Learning rate β_t	Delta content	Control dimension
CUT3R Wang et al. (2025b)	1.0 (constant)	$\mathbf{\Delta}_t$	—
TTT3R Chen et al. (2026)	$\text{sigmoid}(\sum_m \mathbf{Q}_{\mathbf{S}_{t-1}} \mathbf{K}_{\mathbf{X}_t}^\top)$	$\mathbf{\Delta}_t$	Token magnitude
TTSA3R Zheng et al. (2026)	TAUM × SCUM	$\mathbf{\Delta}_t$	Token magnitude
DDD3R (ours)	β_t^{TTT3R}	$\alpha_\perp \boldsymbol{\delta}^\perp + \alpha_\parallel \boldsymbol{\delta}^\parallel$	Token mag. + Direction

A.8 Brake as a Scalar Relaxation of DDD3R

We make precise the relationship between the temporal brake gate $g_t = \text{sigmoid}(-\tau \cos\theta_t)$ with $\cos\theta_t = \cos(\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t-1})$ and the directional decomposition rule $\tilde{\boldsymbol{\delta}}_t = \alpha_\perp \boldsymbol{\delta}_t^\perp + \alpha_\parallel \boldsymbol{\delta}_t^\parallel$ of eq. (6). The two share the same qualitative gating behaviour but differ in expressive power.

Setup. Take the single-frame limit of the EMA tracker, $\mathbf{d}_t = \boldsymbol{\delta}_{t-1}/\|\boldsymbol{\delta}_{t-1}\|$ (i.e. β_{ema}=0). Let θ_t be the angle between $\boldsymbol{\delta}_t$ and $\mathbf{d}_t$. Then $\|\boldsymbol{\delta}_t^\parallel\| = \|\boldsymbol{\delta}_t\|\,|\cos\theta_t|$ and $\|\boldsymbol{\delta}_t^\perp\| = \|\boldsymbol{\delta}_t\|\,|\sin\theta_t|$, with the two components orthogonal.

Effective magnitude of DDD3R. The norm of the regulated delta is

$$\|\tilde{\boldsymbol{\delta}}_t\|^2 = \alpha_\parallel^2\|\boldsymbol{\delta}_t^\parallel\|^2 + \alpha_\perp^2\|\boldsymbol{\delta}_t^\perp\|^2 = \|\boldsymbol{\delta}_t\|^2\big[\alpha_\perp^2 - (\alpha_\perp^2 - \alpha_\parallel^2)\cos^2\theta_t\big], \tag{13}$$

so the *effective scalar gate* that DDD3R applies to $\|\boldsymbol{\delta}_t\|$ is

$$g_t^{\text{DDD3R}}(\cos\theta_t) \;=\; \sqrt{\alpha_\perp^2 - (\alpha_\perp^2 - \alpha_\parallel^2)\cos^2\theta_t}\;. \tag{14}$$

This is a smooth, monotone-decreasing function of $|\cos\theta_t|$, ranging from $\alpha_\perp$ (no drift, $\cos\theta_t{=}0$) to $\alpha_\parallel$ (pure drift, $|\cos\theta_t|{=}1$).

Symmetry comparison with brake. Expanding eq. (14) around $c := \cos\theta_t{=}0$ gives

$$g_t^{\text{DDD3R}}(c) = \alpha_\perp \;-\; \tfrac{\alpha_\perp^2 - \alpha_\parallel^2}{2\alpha_\perp}\,c^2 \;+\; \mathcal{O}(c^4), \tag{15}$$

which is an even function of c: drift along $+\mathbf{d}_t$ and along $-\mathbf{d}_t$ are attenuated identically, because the projection norm $\|\boldsymbol{\delta}_t^\parallel\| = \|\boldsymbol{\delta}_t\|\,|c|$ depends on $|c|$ only. The brake gate $g_t^{\text{brake}}(c) = \text{sigmoid}(-\tau c)$, by contrast, expands as

$$g_t^{\text{brake}}(c) = \tfrac{1}{2} \;-\; \tfrac{\tau}{4}\,c \;+\; \mathcal{O}(c^3), \tag{16}$$

i.e. an *odd* function around $c{=}0$: it dampens only when consecutive deltas are aligned ($c{>}0$) and *passes through or amplifies* when they disagree ($c{<}0$). The two gates therefore agree only on the constant term ($\alpha_\perp{\approx}\frac{1}{2}$) and have qualitatively different behaviour for anti-aligned deltas; brake is not, in any strict Taylor sense, a first-order approximation of DDD3R.

Why brake is still a useful surrogate in practice. Empirically, c is far from zero throughout long sequences: $\bar{c} \in [0.62, 0.77]$ on TUM and ScanNet (table 2). In this practical regime $c{>}0$, both gates are monotone-decreasing in c and pull the update toward a smaller magnitude. Setting $\tau{=}2$ matches their values at the high-drift end ($g^{\text{DDD3R}}(1){=}\alpha_\parallel{=}0.05$ vs. $g^{\text{brake}}(1){=}\text{sigmoid}(-2){\approx}0.12$) and at $c{=}0$ ($\alpha_\perp{=}0.5$ vs. $\text{sigmoid}(0){=}0.5$) to within a factor of $\sim$2. So in the actual operating regime brake delivers *the same qualitative drift–novelty trade-off* as DDD3R, with three caveats below.

Three losses of the relaxation. **(L1) Asymmetric handling of anti-aligned deltas.** For $c{<}0$ (rare but possible during scene transitions), DDD3R still suppresses the drift component to $\alpha_\parallel\|\boldsymbol{\delta}_t\||c|$, while brake *boosts* the update via $\text{sigmoid}(+\tau|c|){>}\frac{1}{2}$. **(L2) No directional separation.** Brake produces a single per-token scalar that multiplies the entire $\boldsymbol{\delta}_t$, so it cannot suppress drift while preserving novelty within the same delta; DDD3R applies $\alpha_\parallel{<}\alpha_\perp$ separately along $\boldsymbol{\delta}_t^\parallel$ and $\boldsymbol{\delta}_t^\perp$. **(L3) Single-frame history.** Brake compares only consecutive deltas, while DDD3R's drift direction is an EMA over the full history; in scenes where underlying drift varies slowly but per-frame deltas are noisy, EMA cancels the noise and brake does not.

When the relaxation is tight. Brake is a near-exact stand-in for DDD3R when (i) $\alpha_\perp{-}\alpha_\parallel$ is small (directional differentiation has little effect), (ii) the per-frame drift direction is stable (single-frame and EMA agree), and (iii) deltas are predominantly aligned (c rarely negative). All three conditions hold in low-drift indoor sequences, which matches our empirical finding that $\text{DDD3R}_{\text{brake}}$ sits very close to $\text{DDD3R}_{\text{ortho}}$ on TUM (gap$\leq$15% relative) and that the gap widens on ScanNet, where drift energy is high and EMA tracking matters more.

Summary. Brake is the *scalar relaxation* of DDD3R: it preserves the qualitative drift-attenuation behaviour in the empirical operating regime $c \in [0.4, 0.8]$, but does not Taylor-match DDD3R near $c{=}0$ (different parity), trades the directional axis for a one-knob per-token gate (L2), and trades the long-history EMA for a single-frame comparison (L3). It is therefore both a strong baseline (often within a few percent of $\text{DDD3R}_{\text{ortho}}$) and a strict subset of eq. (6)'s expressive power.

A.9 Does the Diagnosis Transfer to Another Backbone?

DDD3R regulates the quantity $\boldsymbol{\Delta}_t = \mathbf{S}_t^{\text{new}} - \mathbf{S}_{t-1}$, so it is defined only for models that overwrite a bounded state. Among streaming reconstruction models this is a real restriction. StreamVGGT Zhuo et al. (2026) and Spann3R Wang & Agapito (2025) are append-only, concatenating to a KV cache

or keeping the top-k memory tokens and never editing a token once written, so no $\mathbf{\Delta}_t$ exists there. DUSt3R Wang et al. (2024) and MASt3R Leroy et al. (2024) carry no persistent state at all.

Point3R Wu et al. (2026) does overwrite. It pairs a DUSt3R ViT-L backbone with an explicit spatial pointer memory in which each pointer is anchored to a 3D position and the set grows with the scene, and its merge step replaces a pointer feature outright, which is $\beta = 1$ with no dampening and therefore the extreme of M1. We changed only that assignment to $\mathrm{mem} \leftarrow \mathrm{mem} + \alpha(\mathrm{new} - \mathrm{mem})$, where $\alpha = 1$ reproduces the released model exactly and serves as the baseline.

Table 10: DDD3R's magnitude fix transferred to Point3R (ATE, m ↓). Both datasets select α=0.5, the same coefficient we use on CUT3R.

	upstream (α=1)	α=0.7	α=0.5
ScanNet (n=90)	0.0971	0.0948	**0.0918**
TUM (n=8)	0.0427	0.0382	**0.0341**

M1 transfers. On ScanNet, α=0.5 reduces ATE by 5.4% (Wilcoxon p=0.0022, better on 56/90 scenes). On TUM the effect is larger at 20.1% but n=8 is too small to certify it, so we rely on ScanNet. The effect is real but modest, and it is concentrated in a subset of scenes rather than being a uniform shift.

M3 does not transfer. Point3R's drift energy is 0.164 on the *same* ScanNet scenes where CUT3R's is 0.598, measured over seven sequences across two datasets with a spread of 0.155–0.170, and $\cos(\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t-1})$ is negative. Anchoring each pointer to a 3D position means different viewpoints update it at different times, which decorrelates the update direction. Over-update is thus architecture-independent, while directional redundancy is specific to a bounded, repeatedly overwritten *implicit* state. Point3R escapes that pathology by letting memory grow with the scene, at the cost of bounded memory. This is also what section 3.3 predicts, since low drift energy implies directional decomposition should not be used.

A.10 Real-World Handheld Capture

To test whether the diagnosis holds on data we had never touched, we evaluated on ARKitScenes Baruch et al. (2021), 18 real-world indoor scenes of 500 frames each captured handheld on a consumer iPad Pro, with ground truth from registered laser scans and therefore independent of any visual SfM pipeline. Scenes were fixed in advance by taking the first 20 of the Validation split in ascending video-id order, two of which were too short.

We measured drift energy before running any pose evaluation, since it depends only on the data and the model, obtaining $\bar{e} = 0.632 \pm 0.035$. This is above ScanNet (0.598) and far from TUM (0.398), so section 3.3 predicts that isotropic dampening should win and that directional decomposition should lose, that is, our own default DDD3R$_{\text{ortho}}$ should rank worst. The measured ATE is CUT3R 0.682, TTT3R 0.558, TTSA3R 0.556, DDD3R$_{\text{const}}$ $\mathbf{0.528}$, DDD3R$_{\text{brake}}$ 0.539, DDD3R$_{\text{ortho}}$ 0.609. DDD3R$_{\text{const}}$ improves on CUT3R by 22.6% (p=0.001) and on TTT3R by 5.5% (p=0.027), and DDD3R$_{\text{ortho}}$ is indeed the worst of the three operating points, behind DDD3R$_{\text{const}}$ at p=0.027 on 14/18 scenes.

We also report where this fails. DDD3R$_{\text{ortho}}$ is 9.2% *worse* than TTT3R here, best on only 4/18 scenes. This is the same boundary already visible on ScanNet at 1000 frames, the other high-drift dataset, where DDD3R$_{\text{ortho}}$ (0.488) likewise trails TTT3R (0.406). We therefore scope DDD3R$_{\text{ortho}}$ to low-drift scenes and recommend the isotropic points elsewhere.

A.11 Per-Scene Predictors on ScanNet

Figure 7a tests whether the residual fluctuation of TTT3R's gate carries a timing signal, by running TTT3R on every 1000-frame ScanNet scene, recording the frame-averaged β_t trace and correlating its temporal variance with the per-scene ATE improvement over the un-gated CUT3R baseline. Across n=65 scenes the correlation is at most weakly negative (r=−0.26, p=0.037), and the two split halves give r=−0.05 (p=0.77) and r=−0.42 (p=0.017), so even that marginal significance is unstable.

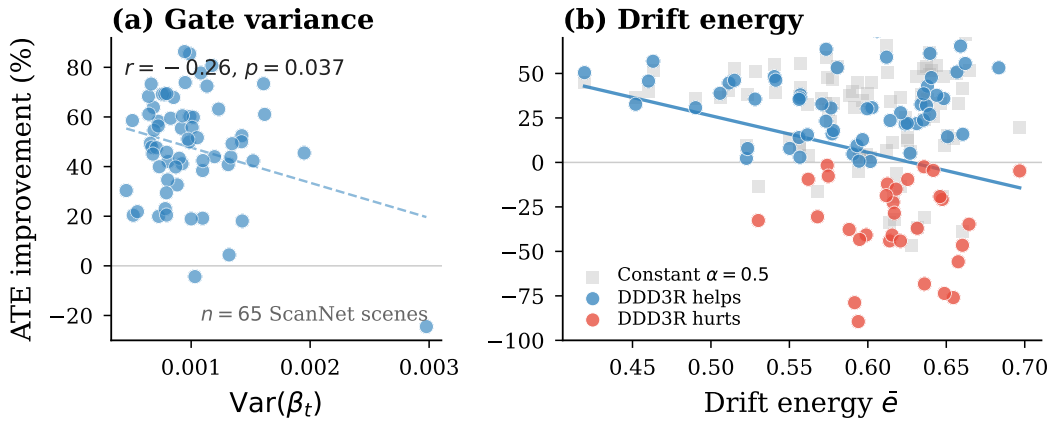

Figure 7: **Per-scene predictors on ScanNet.** **(a)** Direct gate variance $\mathrm{Var}(\beta_t)$ over the full 1000-frame TTT3R trace vs. ATE improvement of TTT3R over the un-gated CUT3R baseline (n=65, 1000-frame valid scenes). The correlation is weak and negative. **(b)** Drift energy $\bar{e}$ vs. ATE improvement of DDD3R over CUT3R (n=90, 90-frame valid scenes). Blue and red dots mark scenes where DDD3R helps or hurts, gray squares are the constant-dampening baseline, and the blue line is the linear fit on DDD3R points.

Higher gate variance corresponds to *less* improvement, the opposite of what an adaptive gate should deliver. Figure 7b is the drift-energy counterpart discussed in section 3.3.