DIFFUSIVE DYNAMICS MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusive Dynamics Modeling uses time as the control variable to predict the diffusion of substances at arbitrary future time points based on current images, which provides a foundation for future generation and real-world modeling. However, this task remains challenging due to: (1) Time-controlled generation of dynamics lacks temporal connections between adjacent time points, leading to discontinuities. (2) Prediction deviations across temporal spans violate the macroscopic directionality, leading to inconsistency. The diffusion model generation process follows diffusive dynamics along the noise-level axis, while real diffusive dynamics evolve along the time axis. Taking these two dynamics as the two axes to form a two-dimensional grid enables the establishment of temporal relationships across multiple noise levels. Therefore, we propose **Diffusive Grid**, a unified diffusive dynamics modeling framework to establish temporal relationships throughout the denoising process. Based on this grid, *Dynamics Alignment Learning (DAL)* aligns predictions that reach the same grid node through temporal evolution and denoising, thereby constraining the entire generation process with adjacent time points and improving continuity. *Grid Walk (GW)* constructs paths across temporal spans and anchors their endpoints to the target, allowing supervision to propagate through intermediate states and regularize consistency. Experiments covering both natural and physiological diffusive processes show that our method achieves better overall performance in continuity and consistency. Code will be available.

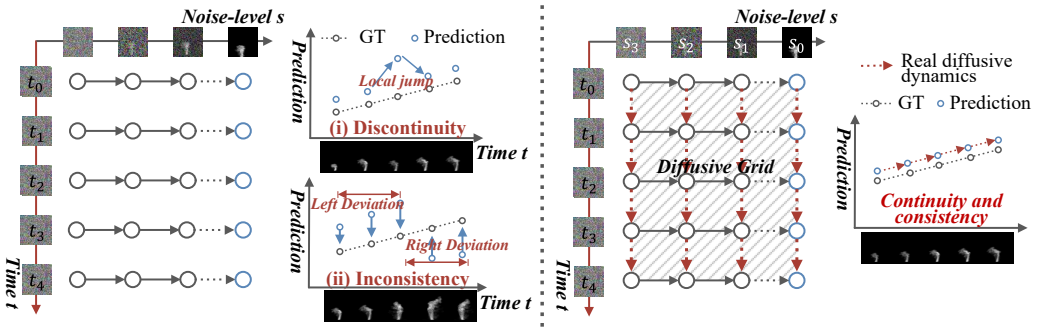
1 INTRODUCTION

Diffusive Dynamics Modeling (DDM) uses time as the control variable to predict the visual observations of diffusive substances at a specified future time based on the current images. Diffusive dynamics has two properties. *(1) Passivity*. Physical or physiological factors drive the evolution of substances and cannot be directly specified through active actions (Warhaft, 2000; Khalifa et al., 2014). Therefore, time is a way to specify their evolution. *(2) Macroscopic directionality.* The overall state of substances exhibits a stable macroscopic evolution, gradually diffusing from local regions to broader areas (Crank, 1979; Kleinfeld, 2002). Such dynamics are widely observed in the real world, such as smoke diffusion in air (Warhaft, 2000) and contrast-agent transport in body tissues (Haddad et al., 2009). Once successfully modeled, it provides a foundation for future generation (Brooks et al., 2024; Dimri et al., 2026) and real-world modeling (Ha & Schmidhuber, 2018).

However, DDM remains challenging due to: (1) *Time-controlled generation of dynamics leads to discontinuity.* Due to the passivity, the predictions are generated at specified times. However, such time-controlled generation overlooks the relationships between neighboring time points, making it difficult for adjacent predictions to preserve the continuous changes inherent in real diffusive dynamics, leading to discontinuity (Guo et al., 2025). (Fig. 1(a.i)). (2) *Prediction deviations across temporal spans lead to inconsistency.* Across temporal spans, prediction deviations at different times accumulate along different directions. When predictions are combined into a temporal trajectory, their direction varies. This violates the macroscopic directionality of the diffusive process, leading to inconsistency (Lai et al., 2018). In Fig. 1(a.ii), the smoke drifts left and then right.

Although existing methods have achieved promising results, they exhibit limitations in DDM. (1) *Temporal regularization methods* (Yang et al., 2023; Liu et al., 2026) for discontinuities constrain adjacent predictions to reduce abrupt changes and improve temporal continuity. However, local constraints suppress the underlying real dynamics, resulting in over-smoothed transitions. (2) *Dynamics-aware generative methods* establish temporal relationships through time-dependent noise

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.



(a) Challenges : ***discontinuity*** from time-controlled generation and ***inconsistency*** from prediction deviations across temporal spans.

(b) Our **Diffusive Grid** jointly models two diffusive dynamics for ***continuity and consistency***

Figure 1: Our Diffusive Grid unifies the generative and real diffusive dynamics to establish temporal connections for continuity and consistency. (a) DDM faces two key challenges, and (b) the Diffusive Grid takes two diffusive dynamics to form a grid and establish the relationships of connecting target-time generations across time and noise levels.

schedules (Ge et al., 2023; Luo et al., 2023) or dependencies on preceding states (Guo et al., 2025) to improve consistency. However, the former rely on predefined noise mechanisms, while the latter embed temporal relations into denoising. Neither directly models real cross-time state transitions, making it difficult to learn the true overall dynamics.

Recently, diffusion models (Ho et al., 2020), which match the diffusive dynamics modeling, have shown promise for future generation. Their generation process is the generative diffusive dynamics along the noise-level axis, while the real diffusive dynamics evolve along the time axis (Guo et al., 2025). If these two dynamics are taken as the two axes to form a two-dimensional grid, it will enable the establishment of temporal relationships across multiple noise levels. For adjacent time points, generation across different noise levels can be connected along the time axis for continuity, while connections across temporal spans further relate predictions at multiple time points for consistency.

In this paper, we propose the ***Diffusive Grid*** for diffusive dynamics modeling, which unifies real and generative diffusive dynamics modeling in a two-dimensional grid to establish temporal connections across target times, thereby improving continuity and consistency. **(1)** ***Dynamics Alignment Learning (DAL) for continuity***: DAL aligns the predictions of the two dynamics at nodes, allowing generation at adjacent times to establish temporal connections across multiple noise levels. Specifically, it matches the predictions reaching the same grid node from the real and generative diffusive dynamics, obtained from the preceding time point and a higher noise level, respectively. By aligning them across noise levels, DAL maintains the temporal connection between adjacent target-time generations throughout denoising, thereby improving continuity. **(2)** ***Grid Walk (GW) for consistency***: GW connects predictions throughout a temporal span to improve consistency. It constructs multiple paths that connect intermediate nodes in the grid and anchors their endpoints to the same ground-truth target. The target supervision acts through each path over the entire temporal span, encouraging generation across multiple time points to maintain a consistent overall direction, thereby improving consistency.

Our contributions are summarized as follows: **(1)** Our proposed **Diffusive Grid** unifies the generative and real diffusive dynamics modeling along the noise-level and time axes, enabling temporal connections across target-time generations at multiple noise levels. **(2)** Our proposed DAL aligns the two dynamics at the same nodes to connect neighboring target-time generations and improve continuity. **(3)** Our proposed Grid Walk (GW) constructs paths in the grid to link predictions across a temporal span for consistency. **(4)** Experiments on both natural and physiological scenarios demonstrate the effectiveness of our method across diverse diffusive dynamics modeling.

2 Related Work

Diffusive Dynamics Modeling. Diffusive dynamics modeling aims to learn the dynamic process by which material distributions continuously diffuse. Its applications involve two types of visual sce-

narios. (1) *Physical scenarios*. Smoke and pollutant dispersion (Warhaft, 2000; Huber et al., 1991) characterize the spatial propagation and concentration evolution of airborne substances, supporting environmental monitoring and dispersion trend prediction. The diffusion of dyes or solutes in fluids (Hirano et al., 2022) characterizes material transport and mixing processes through their evolving spatial distributions. (2) *Physiological scenarios*. Dynamic contrast-enhanced imaging (Tofts et al., 1999; Khalifa et al., 2014) captures the propagation and exchange of contrast agents within tissues to characterize tissue perfusion, providing important information for disease diagnosis and treatment assessment. Fluorescence microscopy characterizes the diffusion and redistribution of fluorescent molecules within cells for analyzing molecular motion and intracellular dynamics (Wachsmuth, 2014; Lorén et al., 2015). These processes exhibit material distributions that continuously evolve, providing important observations for modeling real-world dynamics.

Modeling Time and Noise Level in Diffusion Models. Jointly modeling time t and noise level s enables temporal dynamics to participate in diffusion generation, improving continuity and consistency. (1) *Time-dependent noise modeling* assigns different noise levels to different time points (Ruhe et al., 2024; Chen et al., 2024; Kim et al., 2024; Sun et al., 2025). However, their time–noise correspondence follows predefined schedules, limiting adaptation to diverse real dynamics. (2) *Cross-time state transition modeling* captures temporal evolution through interpolation (Rühling Cachay et al., 2023) or neighboring-state prediction (Guo et al., 2025). However, cross-time evolution is not maintained across noise levels, preventing temporal constraints from persisting throughout generation and causing discontinuity and drift across target times. In contrast, our Diffusive Grid jointly models t and s and coordinates their transitions at grid nodes, propagating cross-time relationships across noise levels for improved continuity and consistency.

Continuity in Prediction. Adjacent-frame continuity requires smooth transitions between neighboring predictions. Independent generation at different target times lacks adjacent-state connections, causing abrupt changes and flickering. Existing methods include: (1) *Smoothness regularization* constrains differences between adjacent latent states (Kong et al., 2026), but strong local constraints may suppress real dynamics and produce rigid or over-smoothed transitions. (2) *Inter-frame relation learning* captures local temporal dependencies (Guo et al., 2023; Ho et al., 2022), yet these relations are learned indirectly through generation objectives, leaving adjacent changes insufficiently constrained. In contrast, DAL explicitly learns cross-time transitions and propagates them across noise levels, preserving adjacent-time relationships throughout generation.

Consistency in Prediction. Consistency requires predictions to maintain coherent direction over long temporal ranges, avoiding accumulated deviations. Existing methods include: (1) *Shared-noise modeling*, which correlates stochastic variables across time through shared or correlated noise (Luo et al., 2023; Ge et al., 2023; Wu et al., 2024). This improves global stability, but shared randomness does not directly constrain how the dynamics evolve. (2) *Shared state modeling*, which reuses reference historical states to preserve information across distant predictions (Guo et al., 2025). However, propagated states maintain content consistency, while deviations in the underlying dynamics still accumulate over time. In Diffusive Grid, GW regularizes temporal evolution through paths across temporal spans, reducing long-range deviation and promoting consistency over time.

3 METHODOLOGY

Our Diffusive Grid (Fig. 2) unifies generative and real diffusive dynamics modeling along the noise-level and time axes to establish temporal connections (Sec. 3.1), aligns the two dynamics at nodes to improve continuity between neighboring predictions (DAL, Sec. 3.2), and constructs paths in the grid to maintain a consistent direction across temporal spans (GW, Sec. 3.3). These components are jointly trained while preserving direct target-time generation at inference (Sec. 3.4).

3.1 DIFFUSIVE GRID FORMULATION

The proposed Diffusive Grid formulates target-time generation as two-dimensional dynamics modeling over time t and noise level s. Within the grid, at a fixed noise level, nodes undergo cross-time state transitions along the time axis, while at a fixed time, nodes are progressively denoised along the noise axis. The two types of transitions intersect at grid nodes, allowing adjacent-time state changes to be associated with the generation process across multiple noise levels, reducing discontinuities

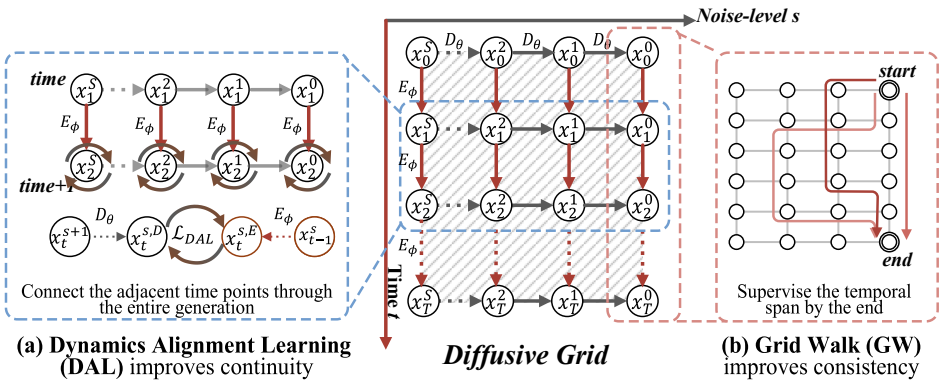


Figure 2: Overview of Diffusive Grid. Diffusive Grid constructs a two-dimensional grid over time t and noise level s, enabling unified modeling of generative and real diffusive dynamics. (a) Dynamics Alignment Learning (DAL) establishes continuous transitions and aligns both at nodes to propagate continuous relationships, promoting continuity. (b) Grid Walk (GW) constrains multiple paths with endpoints to supervise the nodes across the temporal span to enhance consistent direction.

between target times. Furthermore, successive combinations of transitions along the time and noise axes form paths in the grid spanning multiple time points, allowing multiple state changes within a temporal span to be jointly modeled and helping maintain a consistent overall evolution direction.

To characterize the diffusion generation process, we further introduce the noise level $s \in \{0, \dots, S\}$, where $s = 0$ denotes the clean latent and $s = S$ the highest noise level. The latent state at time t and noise level s is defined as $x_t^s = \sqrt{\bar{\alpha}_s} x_t^0 + \sqrt{1 - \bar{\alpha}_s}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. Each state in this diffusive grid is therefore jointly indexed by time t and noise level s, and all such states form the two-dimensional Diffusive Grid:

$$\mathcal{G} = \{x_t^s \mid t \in \mathcal{T},\ s \in \{0, \dots, S\}\}. \tag{1}$$

Within the grid, transitions along the noise axis describe the generative diffusive dynamics, which progressively recover the target state at a fixed time, while transitions along the time axis describe the real diffusive dynamics, which characterize state changes across physical time at a fixed noise level. We parameterize these two dynamics by a Denoiser D_θ and an Evolver E_ϕ, respectively. Their parameters are learned over the Diffusive Grid by

$$(\theta^\star, \phi^\star) = \arg\min_{\theta,\phi} \mathcal{J}_\mathcal{G}(D_\theta, E_\phi), \tag{2}$$

where $\mathcal{J}_\mathcal{G}$ denotes the overall learning objective over the two dynamics and their relations in the grid, which will be specified in the following sections.

3.2 Dynamics Alignment Learning (DAL) Improves Continuity

DAL aligns the predictions of the two dynamics at shared grid nodes, allowing neighboring target-time generations to establish temporal connections across all the noise levels in generation.

As shown in Fig. 2(a), for two adjacent times $t_i < t_{i+1}$ and a noise level s, we regard (t_{i+1}, s) as a shared node that can be reached along either axis of the Diffusive Grid. Along the time axis, the Evolver predicts this node from the preceding state at the same noise level as $\widehat{x}^{s,E}_{t_{i+1}} = E_\phi(x^s_{t_i}, t_i, \Delta t_i, s, c_{t_i})$, where $\Delta t_i = t_{i+1} - t_i$. Along the noise-level axis, the Denoiser reaches the same node from a higher noise level at the target time as $\widehat{x}^{s,D}_{t_{i+1}} = \mathcal{D}_\theta(x^{s+1}_{t_{i+1}}, c_{t_{i+1}})$, where $\mathcal{D}_\theta$ denotes the reverse-diffusion update induced by the Denoiser.

Since the two predictions correspond to the same physical time and noise level, DAL minimizes their discrepancy: $\mathcal{L}_{\text{DAL}} = \mathbb{E}_{t_i,s}\left[\left\|\widehat{x}^{s,E}_{t_{i+1}} - \widehat{x}^{s,D}_{t_{i+1}}\right\|_1\right]$. The alignment loss jointly updates E_ϕ and D_θ, encouraging the two dynamics to produce compatible predictions at shared grid nodes.

Discussion of Continuity. *(1) Adjacent-time Continuity.* E_ϕ models the state transition from t_i to t_{i+1} at a fixed noise level, while D_θ generates the state at t_{i+1} along the noise axis. By jointly op-

timizing the Evolver and Denoiser to align their predictions at the shared node, DAL associates the generation of t_{i+1} with its adjacent-time transition from t_i. *(2) Cross-level Continuity Preservation.* Such alignment is performed across multiple noise levels throughout the whole target-time generation. Therefore, the adjacent-time relation established along the time axis is repeatedly maintained as denoising proceeds along the noise axis. This allows temporal connections between neighboring target times to persist throughout the generation process, improving temporal continuity.

3.3 Grid Walk (GW) Improves Consistency

GW constructs the paths in the grid to connect generation at multiple time points within a temporal span, encouraging a consistent overall evolution direction through target supervision. As shown in Fig. 2(b), given a starting time τ_0 and a target time τ_T randomly selected in a batch, the paths start from the same clean source state $x^0_{\tau_0}$ and eventually reach the level-0 state at the target time τ_T. Multiple paths share the same source-target temporal span, while passing through different noise levels, intermediate time points, and turning positions. The endpoint supervision is therefore able to propagate across the temporal span to improve consistency.

Within each path, the Evolver advances the state to a later time at a fixed noise level, while the Denoiser reduces the noise level at a fixed physical time. Time increases monotonically along the path, whereas the noise level progressively decreases after the initial noising step until the path walking reaches the target node $(\tau_T, 0)$. This construction allows different intermediate state sequences to be formed between the same source and target, thereby involving multiple time points within the temporal span in the same generation path. The detailed algorithm is in Appendix A.2.

Let $\hat{x}^m_{\tau_T}$ denote the prediction reached by the m-th walk at the target node, and let $x^0_{\tau_T}$ denote the ground-truth clean latent at the target time. We anchor the endpoints of M paths to the same ground-truth target: $\mathcal{L}_{\mathrm{GW}} = \frac{1}{M}\sum_{m=1}^{M} \left\| \hat{x}^m_{\tau_T} - x^0_{\tau_T} \right\|_1$.

Discussion of Consistency. GW improves consistency across temporal spans by jointly optimizing multiple temporal transitions under a shared source-to-target objective. Specifically, GW connects intermediate temporal transitions into complete paths and supervises different paths with the same ground-truth endpoint. Since each path has a fixed source, its endpoint error directly reflects the deviation of the accumulated source-to-target change from the ground-truth change. Backpropagating this error through the entire path jointly constrains the temporal transitions toward the same overall evolution, suppressing the accumulation of conflicting directional deviations. During training, randomly sampled source and target times impose such constraints over temporal spans of different lengths and with substantial overlap. As these spans share the same Evolver and Denoiser, individual temporal transitions are jointly constrained by multiple source-to-target objectives, promoting compatible accumulated changes across spans and a consistent macroscopic evolution.

3.4 Training and Inference

The Diffusive Grid is parameterized by the Denoiser D_θ and Evolver E_ϕ. The Denoiser takes the noisy state x^s_t at time t, noise level s, and condition c_t as input, and predicts the diffusion noise as $\widehat{\epsilon}^s_t = D_\theta(x^s_t, s, c_t)$. It is trained with the standard noise-prediction objective $\mathcal{L}_{\mathrm{diff}} = \mathbb{E}[\|D_\theta(x^s_t, s, c_t) - \epsilon\|_2^2]$, establishing the generative diffusive dynamics along the noise-level axis.

The Evolver E_ϕ models state transitions along the time axis using a ResNet-style architecture (Appendix A.1). To learn real diffusive dynamics without interference from random perturbations, $x^s_{t_i}$ and $x^s_{t_{i+1}}$ are constructed at the same noise level using the same Gaussian noise ϵ to construct noisy states at adjacent time points, reducing the influence of random perturbations on temporal transition learning, and the Evolver is trained by $\mathcal{L}_{\mathrm{evo}} = \mathbb{E}[\|E_\phi(x^s_{t_i}, t_i, \Delta t_i, s, c_{t_i}) - x^s_{t_{i+1}}\|_1]$.

Training proceeds in three stages. **Stage 1** separately learns the two basic dynamics. The Denoiser learns the generative diffusive dynamics through noise prediction, while the Evolver learns the real diffusive dynamics through target-state prediction. **Stage 2** further introduces DAL to align the predictions of the two dynamics at shared grid nodes for continuity. **Stage 3** incorporates GW to extend the learning across temporal spans through paths in the grid for consistency. The three training stages are summarized as $\mathcal{L}_1 = \mathcal{L}_{\mathrm{diff}} + \mathcal{L}_{\mathrm{evo}}$, $\mathcal{L}_2 = \mathcal{L}_1 + \lambda_{\mathrm{DAL}}\mathcal{L}_{\mathrm{DAL}}$, $\mathcal{L}_3 = \mathcal{L}_2 + \lambda_{\mathrm{GW}}\mathcal{L}_{\mathrm{GW}}$.

Table 1: Quantitative comparisons on both datasets demonstrate the strong overall performance of Diffusive Grid. The ablation study shows the respective contributions of DAL and GW.

Method	Scalarflow					DCE-MRI				
	MAE↓	PSNR↑	SSIM↑	tRMSE↓	dRMSE↓	MAE↓	PSNR↑	SSIM↑	tRMSE↓	dRMSE↓
Dydiff	$458_{\pm 20.1}$	$16.1_{\pm 0.463}$	$0.697_{\pm 0.0771}$	$0.404_{\pm 0.0403}$	$1.44_{\pm 0.241}$	$290_{\pm 121}$	$19.4_{\pm 3.32}$	$0.296_{\pm 0.126}$	$353_{\pm 153}$	$3.09_{\pm 1.32}$
OU-noise	$97.9_{\pm 13.6}$	$21.2_{\pm 0.948}$	$0.825_{\pm 0.0155}$	$0.258_{\pm 0.0264}$	$0.335_{\pm 0.0404}$	$131_{\pm 82.6}$	$26.0_{\pm 7.73}$	$0.653_{\pm 0.159}$	$209_{\pm 138}$	$1.89_{\pm 1.16}$
STCL	$93.8_{\pm 14.3}$	$21.7_{\pm 1.03}$	$0.831_{\pm 0.0179}$	$0.147_{\pm 0.0251}$	$0.345_{\pm 0.0362}$	$129_{\pm 83.1}$	$26.2_{\pm 7.84}$	$0.660_{\pm 0.160}$	$226_{\pm 153}$	$1.96_{\pm 1.29}$
DDPM	$115_{\pm 11.5}$	$20.3_{\pm 0.642}$	$0.806_{\pm 0.0135}$	$0.322_{\pm 0.0191}$	$0.391_{\pm 0.0325}$	$131_{\pm 82.6}$	$26.0_{\pm 7.71}$	$0.652_{\pm 0.158}$	$226_{\pm 151}$	$1.96_{\pm 1.28}$
+ DAL	$91.5^{*}_{\pm 13.9}$	$\mathbf{22.0}^{*}_{\pm 1.04}$	$0.828^{*}_{\pm 0.0179}$	$0.150^{*}_{\pm 0.0339}$	$0.301^{*}_{\pm 0.0334}$	$129^{*}_{\pm 82.0}$	$26.2^{*}_{\pm 7.83}$	$0.656^{*}_{\pm 0.159}$	$224^{*}_{\pm 149}$	$1.95^{*}_{\pm 1.27}$
Ours(+GW)	$\mathbf{89.9}_{\pm 11.4}$	$22.0_{\pm 1.02}$	$\mathbf{0.843}^{*}_{\pm 0.0196}$	$\mathbf{0.137}^{*}_{\pm 0.0337}$	$\mathbf{0.272}^{*}_{\pm 0.0350}$	$\mathbf{125}^{*}_{\pm 81.9}$	$\mathbf{26.7}^{*}_{\pm 8.12}$	$\mathbf{0.671}^{*}_{\pm 0.168}$	$\mathbf{207}^{*}_{\pm 134}$	$\mathbf{1.79}^{*}_{\pm 1.22}$

Results are reported as mean $\pm$ standard deviation across test cases. Scalarflow MAE values are scaled by 10^3. $*$ denotes a statistically significant improvement over the preceding ablation variant ($p < 0.05$) by a two-sided paired Wilcoxon signed-rank test.

Inference. Given the current observation I_0 and an arbitrary target time t, the model directly generates the future state at the specified target time without sequentially predicting intermediate time points. The target latent is initialized from Gaussian noise and progressively denoised under the conditions of I_0 and t to obtain $\widehat{x}_t^0$, which is then decoded to produce the final prediction $\widehat{I}_t$.

4 Experiments

4.1 Experiment Protocol

Datasets. We evaluate our method on two datasets to examine its generalization across diverse diffusive dynamics in real-world physical and physiological processes. *(1) Scalarflow* (Eckert et al., 2019). Scalarflow is a large-scale real-world smoke dataset that provides temporally evolving observations. It contains complex buoyancy-driven smoke transport and turbulent dynamics, making it well suited for evaluating diffusive dynamics prediction. We use the rendered scalar observations, normalize them to $[-1, 1]$, replicate the single channel to three channels, and resize all frames to 256×256. We split the dataset into training and test sets at the case level with a ratio of 7:3. *(2) Duke Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI)* (Saha et al., 2021). Duke DCE-MRI is a public dataset, which contains dynamic contrast-enhanced scans from 922 patients. The sequential acquisitions capture contrast-agent transport in breast tissue, exhibiting the diffusive dynamics of the contrast agent. Each case contains one pre-contrast and three to four post-contrast acquisitions over roughly 1200 seconds. We split the dataset into training and test sets at the patient level with a ratio of 7:3, the image preprocessing follows (Osuala et al., 2024).

Training Details. (1) Hyper-parameter setting. On Scalarflow, we use a batch size of 10 and train the three stages for a total of 50 epochs. We set $\lambda_{\mathrm{DAL}} = 1.0$ and $\lambda_{\mathrm{GW}} = 0.1$, and use $M = 3$ shared source-target walks in GW. On DCE-MRI, we use a batch size of 4 and train the three stages for a total of 20 epochs, with $\lambda_{\mathrm{DAL}} = 2.0$, $\lambda_{\mathrm{GW}} = 0.1$, and $M = 2$. For both datasets, we use 1,000 diffusion steps and optimize the models with AdamW using a learning rate of 1×10^{-5}. **(2) Experiment Settings.** For *comparison experiments*, we compare against three representative approaches to demonstrate the advantages of our method. (1) Temporal regularization methods: Spatiotemporal Consistency Learning (STCL) (Kong et al., 2026) improves temporal consistency through explicit constraints that regularize variations. (2) Dynamics-aware generative methods: Dynamical Diffusion (Dydiff) (Guo et al., 2025) establishes dependencies on preceding states to jointly model noise-level and time axes. Ornstein-Uhlenbeck correlated noise (OU-noise) (Andrae et al., 2025) promotes temporal coherence by introducing a predefined noise schedule, which enables future prediction at any time without the preceding process. For *ablation experiments*, we adopt DDPM in ControlNet (Zhang et al., 2023) for all variants and progressively introduce DAL and GW to isolate their contributions. **(3) Evaluation Metrics.** Five metrics are employed to evaluate prediction performance from three aspects: *(a) Image Quality.* Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) are used to evaluate the frame-wise fidelity between the prediction $\hat{y}_t$ and the ground-truth y_t at time t. Lower MAE and higher PSNR and SSIM indicate better image quality. *(b) Continuity.* Temporal Root Mean Squared Error (tRMSE) measures adjacent-change errors: $\sqrt{\mathrm{mean}_{t,p}[(\hat{y}_{t+1,p} - \hat{y}_{t,p}) - (y_{t+1,p} - y_{t,p})]^2}$. *(c) Consistency.* Drift Root Mean Squared Error (dRMSE) measures drift relative to the initial frame

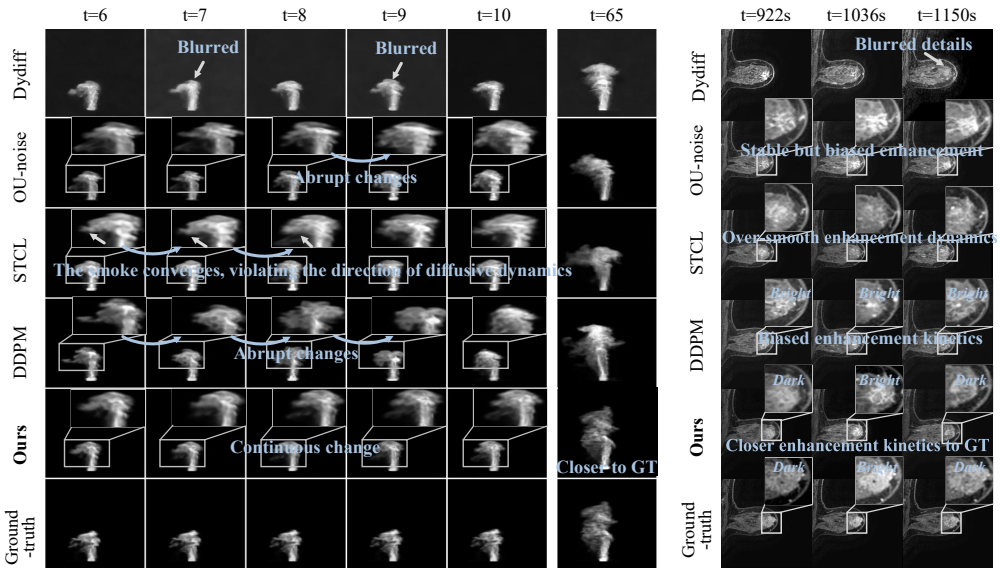


Figure 3: **Visualization** shows that our predictions better follow the ground-truth trend, preserving continuous and consistent smoke dispersion (left) and contrast-agent attenuation (right) trends.

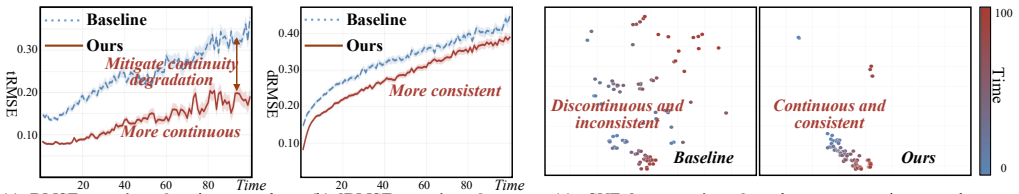
t_1: $\sqrt{\text{mean}_{t>t_1,p}[(\hat{y}_{t,p} - \hat{y}_{t_1,p}) - (y_{t,p} - y_{t_1,p})]^2}$. Here, p indexes spatial locations. Both metrics are computed per sequence and averaged across test cases; lower is better.

4.2 Comparison Analysis

The results in Tab.1 and Fig. 3 demonstrate the advantages of our method in continuity and consistency, with our method achieving almost the best overall performance on both datasets. **(1) Quantitative Analysis.** OU-noise introduces temporally correlated perturbations to improve cross-time prediction. However, such correlation operates at the noise level and does not directly constrain the dynamics between independently generated states, resulting in weaker continuity and consistency than ours. Dydiff jointly models diffusion and temporal evolution but relies on multiple preceding frames; under the single-frame setting of DDM, limited conditioning information makes such dependencies harder to support and increases uncertainty during recursive evolution. STCL substantially reduces tRMSE through smoothness regularization, but its evolution direction is not consistently maintained over time. In contrast, our Diffusive Grid jointly models two diffusive dynamics across time and noise levels, enabling dynamic relations to propagate throughout generation. Consequently, our method achieves nearly all the best scores on both datasets. **(2) Qualitative Analysis.** Our visualization results demonstrate improved temporal continuity and dynamic consistency. As shown in Fig. 3, on Scalarflow (left), our predictions better preserve the continuous upward dispersion of smoke and follow the ground-truth trend over longer target times. In contrast, due to the lack of multiple preceding frames, the Dydiff shows a blurred prediction in Scalarflow. The STCL shows an inconsistent long-range direction, demonstrating the left direction but then right, which is also shown in dRMSE scores. The DDPM shows abrupt changes due to the time-controlled generation of continuous dynamics. On DCE-MRI (right), our method more accurately captures the temporal signal attenuation caused by contrast-agent dispersion, while other methods present biased enhancement or over-smooth dynamics.

4.3 Ablation Study and Model Analysis

Component Ablation. Tab. 1 indicates that our DAL and GW show their contributions on continuity and consistency, respectively. Adding DAL reduces tRMSE on Scalarflow from 0.322 to 0.150, a 53.4% reduction, demonstrating improved continuity between adjacent target times. Further adding GW reduces dRMSE on DCE-MRI from 1.95 to 1.79, an 8.21% reduction, indicating better consistency over temporal spans. The full model achieves the lowest tRMSE and dRMSE on both datasets,



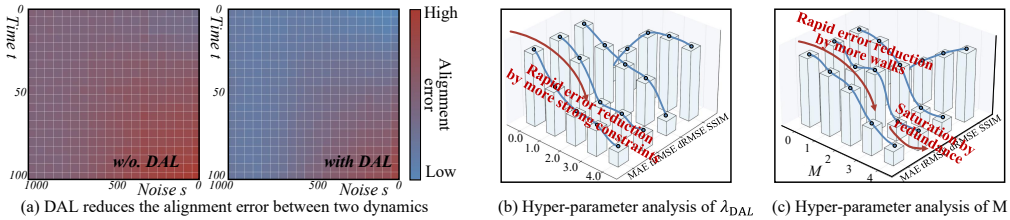
(a) **tRMSE over time**: Ours improves the continuity and mitigates the degradation (b) **dRMSE over time**: Ours shows more consistent dynamics (c) **t-SNE dots over time**: Ours shows more continuous and consistent dots, while baseline shows dispersed dots

Figure 4: **Diffusive Grid promotes continuity and consistency**. (a) Our method achieves lower tRMSE over time, indicating improved temporal continuity and reduced continuity degradation. (b) Ours achieves lower dRMSE over time, indicating more consistent direction from the initial time point. (c) The t-SNE visualization shows a more coherent temporal evolution for our method, while the baseline exhibits more dispersed and irregular state transitions.

confirming the complementary roles of DAL and GW. A two-sided paired Wilcoxon signed-rank test also demonstrates statistically significant improvements over the preceding ablation variants.

Diffusive Grid promotes continuity and consistency. **(1) Analysis of Continuity.** Fig. 4(a) demonstrates the improved temporal continuity of our method. We compute tRMSE at each time point and average it over all test sequences in Scalarflow. Compared with the DDPM baseline, our method consistently achieves lower tRMSE across time, with the error increasing more slowly as the prediction horizon grows. This improvement is attributed to DAL, which aligns the generative and real diffusive dynamics at shared grid nodes and preserves adjacent-time relationships across multiple noise levels throughout denoising. Consequently, the generation at neighboring target times remains better connected, leading to more continuous temporal evolution. **(2) Analysis of Consistency.** Fig. 4(b) demonstrates the improved consistency of our method. We compute dRMSE at each time point and average it over all test cases. Compared with DDPM, our method consistently achieves lower dRMSE over time, indicating reduced accumulated deviation over time. This improvement is attributed to GW, which connects multiple temporal transitions into grid walks and constrains their source-to-target evolution through shared endpoint supervision. By regularizing temporal spans of different lengths, GW suppresses conflicting directional drift across target times and encourages a consistent macroscopic evolution. Consequently, our method better preserves long-range dynamic consistency. **(3) Overall Continuity and Consistency Illustrated by t-SNE.** Fig. 4(c) visualizes the temporal evolution of features in a shared t-SNE space. For the same test case and target times, we jointly project the features of DDPM (left) and Ours (right) into 2D t-SNE space and color each point by time for direct comparison. Compared with DDPM, whose features are more dispersed and follow irregular transitions, our predictions form a more compact and ordered trajectory, with neighboring time points remaining closer and the progression becoming smoother. This behavior reflects the complementary effects of DAL and GW. The DAL preserves adjacent-time relationships across diffusion levels, while the GW regularizes the accumulated evolution across temporal spans. Together, they qualitatively support improved continuity and consistency.

Hyper-parameter Analysis. **(1) The weight λ_{DAL} in DAL.** As shown in Fig. 5(b), performance first improves and then declines as λ_{DAL} increases. A small weight provides insufficient alignment between the Evolver and Denoiser, while a moderate weight effectively coordinates the two dynamics without interfering with their individual supervision. When λ_{DAL} becomes large, the alignment objective dominates optimization and restricts the flexibility of both dynamics, the SSIM decreases slightly. Consequently, the best performance is achieved with a moderate alignment weight $\lambda_{DAL} = 2.0$ in DCE-MRI. **(2) The number of paths M in GW.** As shown in Fig. 5(c), the performance improves because more paths increase path-space coverage, and then saturates because additional paths provide increasingly redundant constraints. We vary the number of stochastic paths M to analyze its effect on dynamic consistency. Increasing M exposes the model to more alternative routes between the same source and target, providing broader sampling of different compositions of temporal evolution and denoising operations. These additional paths provide a more representative approximation of the path-consistency objective and reduce the chance that the model overfits to a limited set of evolution routes. As the major path variations become sufficiently covered, newly sampled paths become increasingly redundant and contribute little additional supervision.

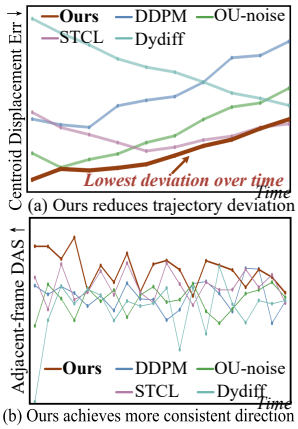


(a) DAL reduces the alignment error between two dynamics (b) Hyper-parameter analysis of λ_{DAL} (c) Hyper-parameter analysis of M

Figure 5: **Each component of Diffusive Grid provides distinct benefits.** (a) DAL aligns the two dynamics to improve continuity. (b) Performance peaks at a moderate λ_{DAL}. (c) Performance improves with M through richer path coverage, then saturates due to redundancy.

DAL Improves Continuity through Dynamics Alignment Learning. Fig. 5(a) shows that DAL substantially reduces the discrepancy between the Evolver and Denoiser at shared grid nodes. For each node (t, s), we compute the L_1 distance between their predictions of the same state and average it over the test set as the node-wise alignment error. Without DAL, clear discrepancies remain across both time and noise levels, whereas DAL consistently lowers the alignment error throughout the grid. This indicates that DAL better coordinates temporal evolution and denoising across noise levels. Together with the reduced tRMSE, these results show that improved dynamics alignment leads to more accurate adjacent-time changes and better temporal continuity.

Diffusive Grid Exhibits Smaller Centroid Trajectory Deviation. As shown in Fig. 6(a), we compute the smoke centroid by weighting pixel locations with smoke density and measure the Euclidean distance between predicted and ground-truth centroids. Our method achieves lower displacement errors over most prediction intervals and a slower error increase with longer horizons. This indicates better tracking of the ground-truth smoke trajectory with reduced positional drift and error accumulation. The improvement arises from constraining temporal state evolution in addition to frame-wise appearance, thereby suppressing deviations over long-term prediction.



(a) Ours reduces trajectory deviation

(b) Ours achieves more consistent direction

Figure 6: Ours shows better performance over time.

Diffusive Grid Better Aligns with the Direction of the Real Dynamics. In Fig. 6(b), we compute the adjacent-frame Directional Alignment Score (DAS) (Eq. 3) at different time points to measure the alignment between the macroscopic changes of adjacent predictions and the ground truth, and compare it with other methods. Our method maintains generally higher and more stable DAS values across time, whereas the baselines exhibit larger fluctuations and more frequent drops in directional alignment. This is because we jointly model generative and real diffusive dynamics, while DAL establishes constraints between adjacent temporal states across different noise levels, allowing temporal transitions to continuously participate in the generation process and reducing stochastic directional deviations from independent generation by GW. Consequently, adjacent predicted states follow the local evolution direction of the real diffusive dynamics more consistently.

5 Conclusion

We propose the Diffusive Grid for diffusive dynamics modeling. It unifies the generative and real diffusive dynamics within a two-dimensional grid, establishing temporal connections among different target times across multiple noise levels. Building on this grid, Dynamics Alignment Learning (DAL) improves local temporal continuity by aligning predictions at the same grid node, while Grid Walk (GW) enhances global evolution consistency through multi-path supervision across temporal spans. Experiments on Scalarflow and DCE-MRI demonstrate the effectiveness of our method in improving prediction continuity and consistency.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for polishing the wording and searching for related work. We have not used generative AI tools for problem formulation, method design, or drawing the plots and figures. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Martin Andrae, Tomas Landelius, Joel Oskarsson, and Fredrik Lindsten. Continuous ensemble weather forecasting with diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 26392–26416, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

John Crank. *The mathematics of diffusion*. Oxford university press, 1979.

Rijul Dimri, Yunsoo Choi, Deveshwar Singh, Jincheol Park, and Nima Khorshidian. Transnet: a transport-informed graph neural network for forecasting pm2. 5 concentrations across south korea. *npj Clean Air*, 2(1):12, 2026.

Marie-Lena Eckert, Kiwon Um, and Nils Thuerey. Scalarflow: a large-scale volumetric data set of real-world scalar transport flows for computer animation and machine learning. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22930–22941, 2023.

Xingzhuo Guo, Yu Zhang, Baixu Chen, Haoran Xu, Jianmin Wang, and Mingsheng Long. Dynamical diffusion: Learning temporal dynamics with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Wassim M Haddad, VijaySekhar Chellaboina, and Sergey G Nersesov. *Thermodynamics: A dynamical systems approach*. Princeton University Press, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hiroyuki Hirano, Shogo Tsuzaki, Hideaki Obata, and Tomohiko Kihara. A colorimetric method for quantitative visualization of diffusion and internal circulation in liquid–liquid two-phase flow. *Chemical Engineering Science*, 249:117285, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Alan H Huber, S Pal Arya, Sarah A Rajala, and James W Borek. Preliminary studies of video images of smoke dispersion in the near wake of a model building. *Atmospheric Environment. Part A. General Topics*, 25(7):1199–1209, 1991.

Fahmi Khalifa, Ahmed Soliman, Ayman El-Baz, Mohamed Abou El-Ghar, Tarek El-Diasty, Georgy Gimel'farb, Rosemary Ouseph, and Amy C Dwyer. Models and methods for analyzing dce-mri: A review. *Medical physics*, 41(12):124301, 2014.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37: 89834–89868, 2024.

David Kleinfeld. Diffusive dynamics. Lecture notes for Physics 171/271, 2002. URL `https://neurophysics.ucsd.edu/courses/physics_171/chapt_3.pdf`. Fall 2002 semester.

Jindi Kong, Yuting He, Cong Xia, Rongjun Ge, and Shuo Li. Mri contrast enhancement kinetics world model. *arXiv preprint arXiv:2602.19285*, 2026.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Enhuai Liu, Yunke Wang, Changming Sun, and Chang Xu. D2cache: Second-order delta caching for higher video diffusion acceleration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43589–43599, 2026.

Niklas Lorén, Joel Hagman, Jenny K Jonasson, Hendrik Deschout, Diana Bernin, Francesca Cella-Zanacchi, Alberto Diaspro, James G McNally, Marcel Ameloot, Nick Smisdom, et al. Fluorescence recovery after photobleaching in material and life sciences: putting theory into practice. *Quarterly reviews of biophysics*, 48(3):323–387, 2015.

Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023.

Richard Osuala, Daniel M Lang, Preeti Verma, Smriti Joshi, Apostolia Tsirikoglou, Grzegorz Skorupko, Kaisar Kushibar, Lidia Garrucho, Walter HL Pinaya, Oliver Diaz, et al. Towards learning contrast kinetics with multi-condition latent diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 713–723. Springer, 2024.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42818–42835. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/ruhe24a.html`.

Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *Advances in neural information processing systems*, 36:45259–45287, 2023.

Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Jingxi Weng, Elizabeth H Cain, Connie E Kim, Sujata V Ghate, Ryan Walsh, and Maciej A Mazurowski. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. *The Cancer Imaging Archive*, 10:5, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, SiYu Zhou, Qian He, and Jing Liu. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7364–7373. IEEE, 2025.

Paul S Tofts, Gunnar Brix, David L Buckley, Jeffrey L Evelhoch, Elizabeth Henderson, Michael V Knopp, Henrik BW Larsson, Ting-Yim Lee, Nina A Mayr, Geoffrey JM Parker, et al. Estimating kinetic parameters from dynamic contrast-enhanced t1-weighted mri of a diffusable tracer: standardized quantities and symbols. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 10(3):223–232, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Malte Wachsmuth. Molecular diffusion and binding analyzed with frap: M. wachsmuth. *Protoplasma*, 251(2):373–382, 2014.

Zellman Warhaft. Passive scalars in turbulent flows. *Annual review of fluid mechanics*, 32(1):203–240, 2000.

Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot video generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7089–7098. IEEE, 2024.

Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14592–14601. IEEE, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

A METHOD DETAILS

A.1 THE ARCHITECTURE OF EVOLVER

Evolver Architecture. The Evolver E_ϕ models state transitions along the time axis t at a fixed noise level s. Given a source latent state $x^s_{t_i}$, source time t_i, time interval $\Delta t_i = t_{i+1} - t_i$, noise level s, and condition c_{t_i}, it directly predicts the target latent state at the same noise level: $\widehat{x}^s_{t_{i+1}} = E_\phi\left(x^s_{t_i}, t_i, \Delta t_i, s, c_{t_i}\right)$. Here, c_{t_i} includes the spatial condition (the initial image) and the text condition (the target time) used by the Evolver.

Condition Injection. The Evolver further conditions the state transition on the source time t_i, time interval Δt_i, and noise level s. These three scalars are separately encoded using 32-dimensional sinusoidal embeddings, concatenated, and projected by an MLP into a 128-dimensional condition representation. The text condition is mean-pooled and projected by another MLP into the same feature dimension, and is then added to the temporal condition. The resulting global condition is injected into all eight residual blocks through FiLM (Perez et al., 2018), which generates channel-wise scale and shift parameters for feature modulation. In parallel, the spatial condition is directly concatenated with the source latent at the network input.

Residual Backbone. The Evolver adopts a resolution-preserving ResNet-style backbone (He et al., 2016). Specifically, the 4-channel source latent and 4-channel spatial condition are concatenated into an 8-channel input and projected to 128 feature channels by a 3×3 convolution. The features are then processed by eight conditional residual blocks, all maintaining 128 channels. Each block contains two 3×3 dilated convolutions and an identity residual connection. The dilation rates are set to $(1, 2, 4, 8, 8, 4, 2, 1)$, which progressively enlarge the receptive field without reducing the spatial resolution. After the residual backbone, Group Normalization and SiLU activation are followed by a final 3×3 convolution that maps the 128-channel representation back to four channels, directly producing the target latent $\widehat{x}^s_{t_{i+1}}$.

A.2 THE DETAILS OF GRID WALK

As shown in Tab. 2, Q_s denotes a single forward-noising operation from the clean state to noise level s, while $D^{(N)}_\theta$ denotes a deterministic DDIM transition from a higher noise level to a lower noise level using N denoising updates (Song et al., 2020). All paths share the same ground-truth source state and temporal interval.

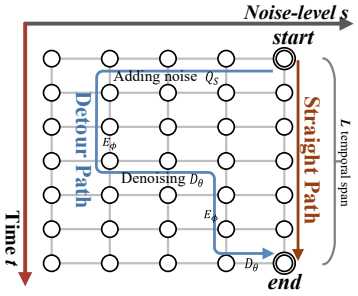


Figure 7: The details of the Grid Walk operation.

For the m-th detour path, J_m denotes the number of denoising segments, $\{r_{m,j}\}_{j=0}^{J_m}$ specifies the temporal turning positions, and $\{s_{m,j}\}_{j=0}^{J_m}$ specifies the corresponding monotonically decreasing noise levels. The path is forward-noised only once at the starting position on the selected time t in a batch. Within each temporal segment $[t_{r_{m,j}}, t_{r_{m,j+1}}]$, the Evolver advances the state autoregressively along the time axis while keeping the noise level $s_{m,j}$ fixed. At the end of the segment, the Denoiser moves the state from $s_{m,j}$ to $s_{m,j+1}$ while keeping the physical time fixed. Therefore, both E_ϕ and D_θ can be invoked multiple times within a single source-to-target path, allowing the path to alternate repeatedly between the time and noise-level axes.

Each path invokes the Evolver L times in total, where the prediction from the previous step is directly used as the input to the next step without replacing intermediate states with ground truth. The final noise level is fixed as $s_{m,J_m} = 0$, ensuring that every detour path terminates at the clean target node.

Table 2: **Grid Walk operation.** H denotes the temporal-window horizon, measured by the number of adjacent transitions, and L denotes the length of a temporal span sampled within that window. M denotes the total number of paths, and $\{\tau_0, \ldots, \tau_L\}$ denotes the sampled temporal span, where τ_0 and τ_L are its source and target times. Each detour path may alternate multiple times between temporal evolution and denoising.

Step	Operation
Input	Temporal-window horizon range $[H_{\min}, H_{\max}]$, temporal-span length range $[L_{\min}, L_{\max}]$, total number of paths M, high-noise-level range $[s_{\min}, s_{\max}]$, and number of denoising steps N.
1	Uniformly sample an integer horizon $H \in [H_{\min}, H_{\max}]$, and randomly select a contiguous temporal window containing H adjacent transitions: $\{(x^0_{t_q}, t_q, c_{t_q})\}_{q=0}^{H}$.
2	Uniformly sample an integer span length $L \in [L_{\min}, \min(L_{\max}, H)]$ and a start position $a \in \{0, \ldots, H-L\}$. Set $\tau_i = t_{a+i}$ for $i = 0, \ldots, L$, yielding the shared span $\{(x^0_{\tau_i}, \tau_i, c_{\tau_i})\}_{i=0}^{L}$.
3	Initialize the straight-path state as $u \leftarrow x^0_{\tau_0}$.
4	**for** $i = 0, \ldots, L-1$ **do**
5	$u \leftarrow E_\phi\big(u, \tau_i, \tau_{i+1} - \tau_i, 0, c_{\tau_i}\big)$.
6	**end for**; save the straight-path endpoint as $\widehat{x}^0_{\tau_L,0} \leftarrow u$.
7	**for** $m = 1, \ldots, M-1$ **do**
8	Sample the number of denoising segments $J_m \in \{1, \ldots, L\}$, and sample temporal turning positions $0 = r_{m,0} < r_{m,1} < \cdots < r_{m,J_m} = L$.
9	Sample a monotonically decreasing noise-level sequence $s_{m,0} > s_{m,1} > \cdots > s_{m,J_m} = 0$, where $s_{m,0} \sim \mathcal{U}[s_{\min}, s_{\max}]$.
10	Sample $\epsilon_m \sim \mathcal{N}(0, I)$ and initialize the detour state as $z \leftarrow Q_{s_{m,0}}\big(x^0_{\tau_0}, \epsilon_m\big)$.
11	**for** $j = 0, \ldots, J_m - 1$ **do**
12	**for** $i = r_{m,j}, \ldots, r_{m,j+1} - 1$ **do**
13	$z \leftarrow E_\phi\big(z, \tau_i, \tau_{i+1} - \tau_i, s_{m,j}, c_{\tau_i}\big)$.
14	**end for**
15	At the fixed physical time $\tau_{r_{m,j+1}}$, denoise the state from $s_{m,j}$ to $s_{m,j+1}$: $z \leftarrow D^{(N)}_\theta\big(z, s_{m,j} \to s_{m,j+1}, c_{\tau_{r_{m,j+1}}}\big)$.
16	**end for**
17	Save the endpoint as $\widehat{x}^0_{\tau_L,m} \leftarrow z$.
18	**end for**
Output	M path endpoints $\{\widehat{x}^0_{\tau_L,m}\}_{m=0}^{M-1}$.

B EXPERIMENTS

B.1 TRAINING SETUP DETAILS

On the Scalarflow dataset, we set the loss weight for Dynamics Alignment Learning to $\lambda_{\mathrm{DAL}} = 1.0$. For Grid Walk, the temporal window length is set to $H = 4$, yielding $H + 1$ contiguous observations from the same trajectory. The minimum and maximum detour lengths are set to $L_{min} = 1$ and $L_{max} = 4$, respectively. The total number of paths in each temporal window is set to $M = 3$, including one straight path and 2 detour paths. The high-noise-level sampling range is set to $[s_{\min}, s_{\max}] = [50, 999]$. For each detour path, the initial noise level is sampled as $s_{m,0} \sim \mathcal{U}[s_{\min}, s_{\max}]$, and the Gaussian noise is independently sampled as $\epsilon_m \sim \mathcal{N}(0, I)$. The corresponding temporal turning positions are denoted by $0 = r_{m,0} < r_{m,1} < \cdots < r_{m,J_m} = L$, and the monotonically decreasing noise-level sequence is denoted by $s_{m,0} > s_{m,1} > \cdots > s_{m,J_m} = 0$. At each turning position, the state is denoised from $s_{m,j}$ to $s_{m,j+1}$ using $N = 4$ deterministic DDIM updates. The DDIM stochasticity parameter is set to $\eta = 0$. The diffusion process contains $S = 1000$ diffusion steps. The batch size is set to $B = 10$, and the optimizer learning rate is set to $\mathrm{lr} = 0.00001$. The first, second, and third training stages are conducted for $E_1 = 25$, $E_2 = 15$, and $E_3 = 10$ epochs.

In the DCE-MRI dataset, we set $\lambda_{DAL} = 2.0$ for dynamics alignment learning. For Grid Walk, each window uses $M = 2$ paths, comprising one straight path and one detour path, with a fixed base noise level of 0. The integer high-noise sampling range is $[1, 99]$ for more detailed generation. The

Table 3: Quantitative comparisons on the Scalarflow dataset at *unseen* future time points.

Method	Scalarflow				
	MAE↓	PSNR↑	SSIM↑	tRMSE↓	dRMSE↓
Dydiff	$642_{\pm 35.4}$	$13.6_{\pm 0.504}$	$0.0802_{\pm 0.0104}$	$0.405_{\pm 0.0497}$	$0.808_{\pm 0.160}$
OU-noise	$234_{\pm 47.9}$	$17.9_{\pm 1.14}$	$0.192_{\pm 0.0768}$	$\mathbf{0.0785}^{\dagger}_{\pm 0.0112}$	$\mathbf{0.412}^{\dagger}_{\pm 0.0489}$
STCL	$160_{\pm 25.3}$	$17.9_{\pm 1.09}$	$0.665_{\pm 0.0296}$	$0.190_{\pm 0.0375}$	$0.475_{\pm 0.0838}$
DDPM	$153_{\pm 25.0}$	$18.0_{\pm 1.10}$	$0.735_{\pm 0.0225}$	$0.175_{\pm 0.0362}$	$0.432_{\pm 0.0614}$
Ours	$\mathbf{147}_{\pm 25.9}$	$\mathbf{18.3}_{\pm 1.17}$	$\mathbf{0.759}_{\pm 0.0202}$	$0.149_{\pm 0.0363}$	$0.417_{\pm 0.0660}$

Scalarflow MAE values (including standard deviations) are scaled by 10^3. Results are reported as mean $\pm$ standard deviation across test cases.
† Lowest temporal error, but accompanied by substantially degraded frame-level fidelity shown in MAE, PSNR and SSIM.

dataset is also split 7:3 for training and testing. The batch size is set to $B = 4$, and the optimizer learning rate is set to $\mathrm{lr} = 0.00001$. The first, second, and third training stages are conducted for $E_1 = 10$, $E_2 = 6$, and $E_3 = 4$ epochs. Unless otherwise specified, the remaining Grid Walk and optimization settings follow those used for Scalarflow.

Fixed sine-cosine positional encoding Vaswani et al. (2017) is adopted for time t to accommodate predictions at arbitrary time instants. The Denoiser uses a U-Net following (Zhang et al., 2023).

B.2 Dataset Details

Scalarflow Dataset. We use the rendered density images from the Scalarflow dataset (Eckert et al., 2019). Scalarflow records real smoke plumes using five synchronized and calibrated cameras at 60 frames per second. We use the first calibrated camera view (`view 0`). Each original rendered sample contains five single-channel views with a spatial resolution of 600×1062 pixels. Before being fed into the network, the selected view is normalized to $[-1, 1]$, and replicated across three channels. For each sequence, we retain the consecutive frames from frame 50 to frame 150, inclusive, and use frame 50 as the initial conditioning state. By frame 50, a visible and spatially structured plume has formed in all sequences, providing a non-trivial initial condition while still preserving substantial subsequent dynamics. One temporal unit in our implementation corresponds to approximately 16.67 ms of physical time. **DCE-MRI Dataset.** We use the Duke-Breast-Cancer-MRI dataset (Saha et al., 2021), a public pre-operative breast MRI dataset containing 922 patients with biopsy-confirmed invasive breast cancer. Each examination includes a pre-contrast acquisition and typically three to four post-contrast acquisitions, capturing the temporal enhancement and redistribution of contrast agent in breast tissue. In our experiments, each axial slice is treated as a 2D sample, resulting in 80,548 slices in total, with the pre-contrast image used as the initial condition and the actual acquisition times of the post-contrast scans used as target times.

B.3 Comparison analysis

Comparison Experiments at arbitrary future time points. In Tab. 3, Diffusive Grid generates accurate predictions at arbitrary time points, including unseen time points, while maintaining competitive continuity and consistency. To evaluate generalization beyond the observed temporal range, we train all methods using target times from only the first 70% of the temporal range and evaluate them on the remaining 30% using disjoint test cases. Therefore, both the test cases and the evaluated future target times are unseen during training. Diffusive Grid achieves the best frame-level prediction accuracy under this challenging extrapolation setting, obtaining an MAE of 147, a PSNR of 18.3, and an SSIM of 0.759. Compared with DDPM, our method reduces tRMSE from 0.175 to 0.149 and dRMSE from 0.432 to 0.417, while simultaneously improving all three fidelity metrics. Compared with STCL, Diffusive Grid further reduces tRMSE by 21.6% and dRMSE by 12.2%, demonstrating improved temporal prediction beyond the target-time range observed during training. Although OU-noise achieves a lower tRMSE and a slightly lower dRMSE, an MAE of 234 and an SSIM of 0.192 demonstrate that it substantially degrades frame-level fidelity. Dydiff shows pronounced degradation at unseen future target times, with substantially higher MAE, tRMSE, and dRMSE. Overall, these results demonstrate that the temporal relationships learned by Diffusive Grid remain effective when extrapolating to unseen cases outside the training range.

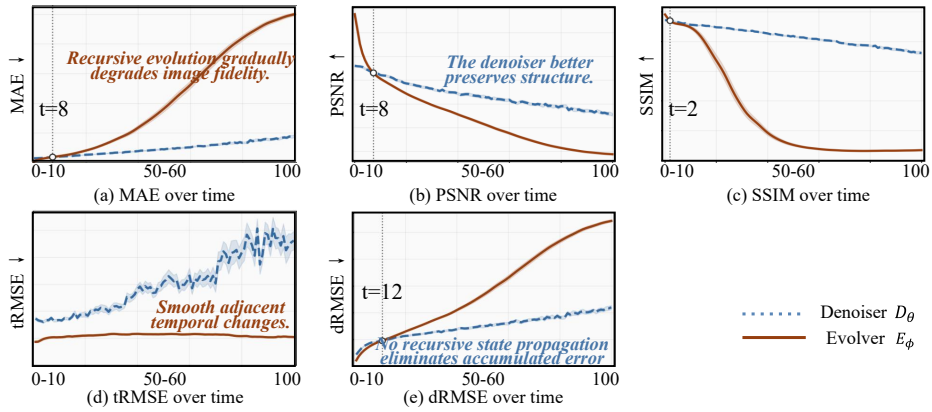


Figure 8: Comparison of recursive Evolver and direct Denoiser inference. Evolver yields smoother adjacent-time evolution but suffers from accumulated errors during recursive rollout, whereas Denoiser avoids recursive propagation and better preserves image fidelity and long-term consistency.

B.4 More model analysis

Comparison between Evolver and Denoiser. Since our Diffusive Grid jointly learns generative and real diffusive dynamics, we further analyze their behavior as alternative inference paths. As shown in Fig. 8, we compare recursive evolution along the temporal axis using the Evolver with direct generation at specified target times using the Denoiser. The Evolver consistently achieves lower tRMSE, indicating that step-by-step transitions better preserve adjacent-time changes. However, as recursive steps increase, its MAE and dRMSE increase, while PSNR and SSIM decrease. Recursive inference feeds previous predictions into subsequent transitions, allowing local errors to accumulate and induce state-distribution shift and long-term trajectory drift. In contrast, the Denoiser generates each target state directly, avoiding recursive error accumulation and maintaining more stable image fidelity and long-term evolution. Although direct generation yields relatively higher tRMSE, joint training constrains the Denoiser through the learned real diffusive dynamics, allowing its predictions to retain cross-time dynamical information. These results show that joint learning allows the Evolver to regularize generation while enabling direct Denoiser-based inference to maintain favorable continuity and long-term consistency without recursive error accumulation.

Evolver Analysis. We additionally perform Evolver rollout on the *initial noisy states*, followed by standard denoising, to evaluate the learned temporal dynamics. As shown in Tab. 4, on DCE-MRI, this rollout improves MAE, PSNR, and SSIM, while reducing tRMSE and dRMSE by 41.71 and 0.3663, respectively. By learning cross-time state transitions at the same noise level, the Evolver introduces target-time-related dynamic information into noisy states before denoising, improving subsequent generation. These gains support the predictive value of the learned temporal transitions. This rollout is used only to analyze the Evolver; the main inference procedure still uses direct target-time generation without sequential rollout. Our method also outperforms OU-noise, which uses a fixed-form noise rollout strategy, suggesting that learned temporal transitions provide more effective target-time dynamic information than predefined noise.

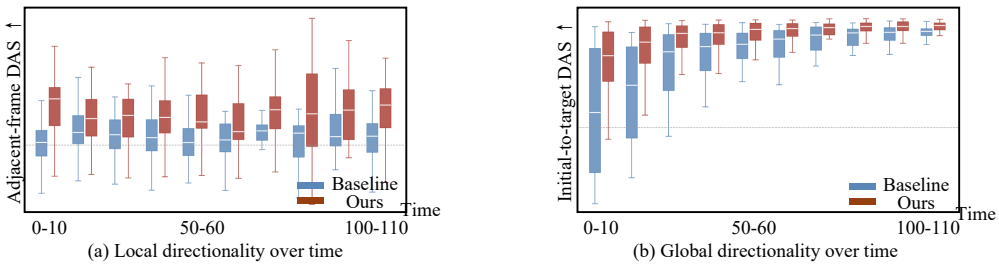
Table 4: Improvement after utilizing Evolver to roll out from the initial noisy state to generate.

MAE ↓	PSNR ↑	SSIM ↑	tRMSE ↓	dRMSE ↓
−4.354	+0.3382	+0.0116	−41.71	−0.3663

Different Types of Path. Detour Path jointly exploits state evolution along the time axis and generative recovery along the noise axis, providing additional dynamic constraints for target-time prediction and thereby improving temporal continuity. To verify this effect, we compare Straight Path and Detour Path in the Scalarflow dataset. Straight Path advances to

Table 5: Comparison between straight and detour paths.

Walk Types	MAE ↓	PSNR ↑	SSIM ↑	tRMSE ↓	dRMSE ↓
Straight Path	**91.50**	21.98	**0.8282**	0.1495	0.3010
Detour Path	92.41	**22.04**	0.8239	**0.1399**	**0.3009**



(a) Local directionality over time (b) Global directionality over time

Figure 9: Additional directionality analysis of GW. (a) Local directionality measured by adjacent-frame Macro-DAS. (b) Global directionality measured by initial-to-target Macro-DAS.

the target time solely through the Evolver along
the time axis, whereas Detour Path first perturbs the state to a higher noise level, performs temporal evolution with the Evolver at that noise level, and then recovers the clean target state through the Denoiser. As shown in Tab. 5, Detour Path reduces tRMSE from 0.1495 to 0.1399, while also achieving higher PSNR and a slightly lower dRMSE. This is because Straight Path mainly relies on the temporal transitions learned by the Evolver. In contrast, Detour Path further introduces the generative diffusion prior learned by the Denoiser, allowing the target state to be jointly shaped by temporal dynamics and generative recovery. This reduces cross-time prediction errors and improves continuity. Meanwhile, Straight Path performs slightly better in MAE and SSIM, indicating that the two paths provide complementary benefits. Therefore, we adopt one Straight Path together with Detour Paths in Grid Walk to retain the complementary advantages of direct temporal evolution and detour-based generative recovery.

Additional Analysis of Continuity. Our method produces more continuous predictions across neighboring target times. Given a predicted sequence $\{\hat{I}_t\}_{t=1}^{T}$, we compute the consecutive structural similarity as $\text{cSSIM} = \frac{1}{T-1}\sum_{t=1}^{T-1}\text{SSIM}(\hat{I}_t, \hat{I}_{t+1})$, where a higher value indicates fewer abrupt structural changes between adjacent predictions. On Scalarflow, densely sampled ground-truth sequences are available. We therefore additionally report $\text{cSSIM-err} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left|\text{SSIM}(\hat{I}_t, \hat{I}_{t+1}) - \text{SSIM}(I_t, I_{t+1})\right|$, which measures whether the structural variation between predicted adjacent frames matches that of the ground truth.

As shown in Tab. 6, our method reduces cSSIM-err from 0.1215 to 0.0341, corresponding to a 71.9% reduction, while increasing cSSIM from 0.8553 to 0.9430. These results indicate that the improved continuity is not merely caused by making adjacent predictions more similar; rather, the magnitude of their structural changes becomes substantially closer to the real temporal evolution. For DCE-MRI, densely sampled ground-truth frames between consecutive acquisitions are unavailable. We therefore report cSSIM only as a no-reference measure of temporal regularity. Our method improves cSSIM from 0.6113 to 0.6609, indicating smoother structural evolution across queried target times. This result should be interpreted together with the improved frame-level fidelity reported in Tab. 1, which prevents the higher cSSIM from being simply caused by static predictions. Diffusive Grid aligns temporal evolution and generative denoising at nodes across multiple noise levels. Consequently, the adjacent-time relationships learned by the Evolver regularize the Denoiser throughout generation, suppressing abrupt temporal fluctuations and producing more continuous predictions.

Table 6: Additional continuity comparison on both datasets.

Method	Scalarflow		DCE-MRI
	cSSIM-err ↓	cSSIM ↑	cSSIM ↑
Baseline	0.1215	0.8553	0.6113
Ours	**0.0341**	**0.9430**	**0.6609**

Additional Analysis of Directional Consistency. As shown in Fig. 9, our method exhibits better macroscopic directional consistency than the baseline in both local and global temporal evolution. Specifically, we represent each frame using a macroscopic state vector $\mathbf{z}_t = [c_x(t), c_y(t), s_x(t), s_y(t)]$, where c_x and c_y denote the density-weighted center-of-mass coordinates, while s_x and s_y characterize the spatial spread of the density field along the two axes. We then quantify the alignment between the predicted and ground-truth macroscopic changes using the macro-

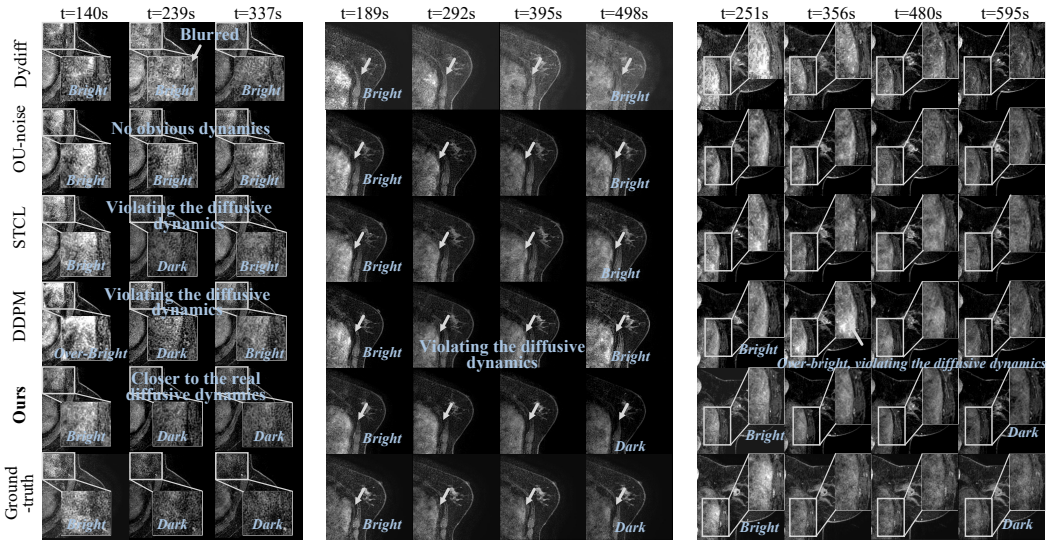


Figure 10: **Visualization** in DCE-MRI shows that our predictions better follow the ground-truth trend. As contrast agent diffuses within tissues, our method better captures the bright-to-dark trend.

scopic Directional Alignment Score (DAS):

$$\mathrm{DAS}\left(\Delta\widetilde{\mathbf{z}}^{\mathrm{pred}}, \Delta\widetilde{\mathbf{z}}^{\mathrm{GT}}\right) = \frac{\Delta\widetilde{\mathbf{z}}^{\mathrm{pred}} \cdot \Delta\widetilde{\mathbf{z}}^{\mathrm{GT}}}{\|\Delta\widetilde{\mathbf{z}}^{\mathrm{pred}}\|_2 \, \|\Delta\widetilde{\mathbf{z}}^{\mathrm{GT}}\|_2 + \epsilon}, \tag{3}$$

where each feature dimension is normalized by its standard deviation $\boldsymbol{\sigma}$ computed from the ground-truth data. For the local directionality analysis in Fig. 9(a), we use the adjacent-frame change $\Delta\widetilde{\mathbf{z}}_t = (\mathbf{z}_t - \mathbf{z}_{t-1})/\boldsymbol{\sigma}$ to measure directional consistency over short time scales. For the global directionality analysis in Fig. 9(b), we instead use the cumulative net change from the initial frame to each target frame, $\Delta\widetilde{\mathbf{z}}_{0\to t} = (\mathbf{z}_t - \mathbf{z}_0)/\boldsymbol{\sigma}$, to assess long-term trajectory alignment between the prediction and ground truth. For visualization, we first average the DAS values for each sequence within each time interval and then use box plots to show their distributions across sequences. A higher DAS indicates that the overall displacement and deformation of the predicted density field are more closely aligned with the ground-truth evolution. Our method produces consistently higher DAS distributions than the baseline for both adjacent-frame and initial-to-target comparisons, demonstrating more accurate evolution and better preservation of long-term trajectory direction. This improvement stems from GW's explicit modeling of cross-time evolution: it promotes consistent transitions between neighboring states and suppresses inconsistent directional deviations, enabling the predicted dynamics to follow the ground-truth trajectory more faithfully.

B.5 More visualization

Our method demonstrates better continuity and consistency across multiple examples. As shown in Fig. 10 and Fig. 11, existing methods exhibit noticeable discontinuity and inconsistency during temporal evolution. For local continuity, DDPM shows abrupt intensity changes between adjacent time points, while OU-noise and STCL also exhibit abrupt changes at later stages, indicating unstable transitions between neighboring states. For long-span consistency, Dydiff exhibits abrupt direction changes, while DDPM evolves toward various directions. OU-noise shows different directions over time in the second example and even violates the diffusive dynamics in the third example, where the smoke shrinks against the real dispersion trend. In contrast, our Diffusive Grid jointly models real and generative diffusive dynamics: DAL establishes connections between adjacent time points across different noise levels to preserve local continuous changes, while GW further constrains the overall evolution across temporal spans. Consequently, our predictions maintain smoother state changes between adjacent times and consistently follow the overall ground-truth dispersion trend over long temporal spans, demonstrating improved continuity and consistency.

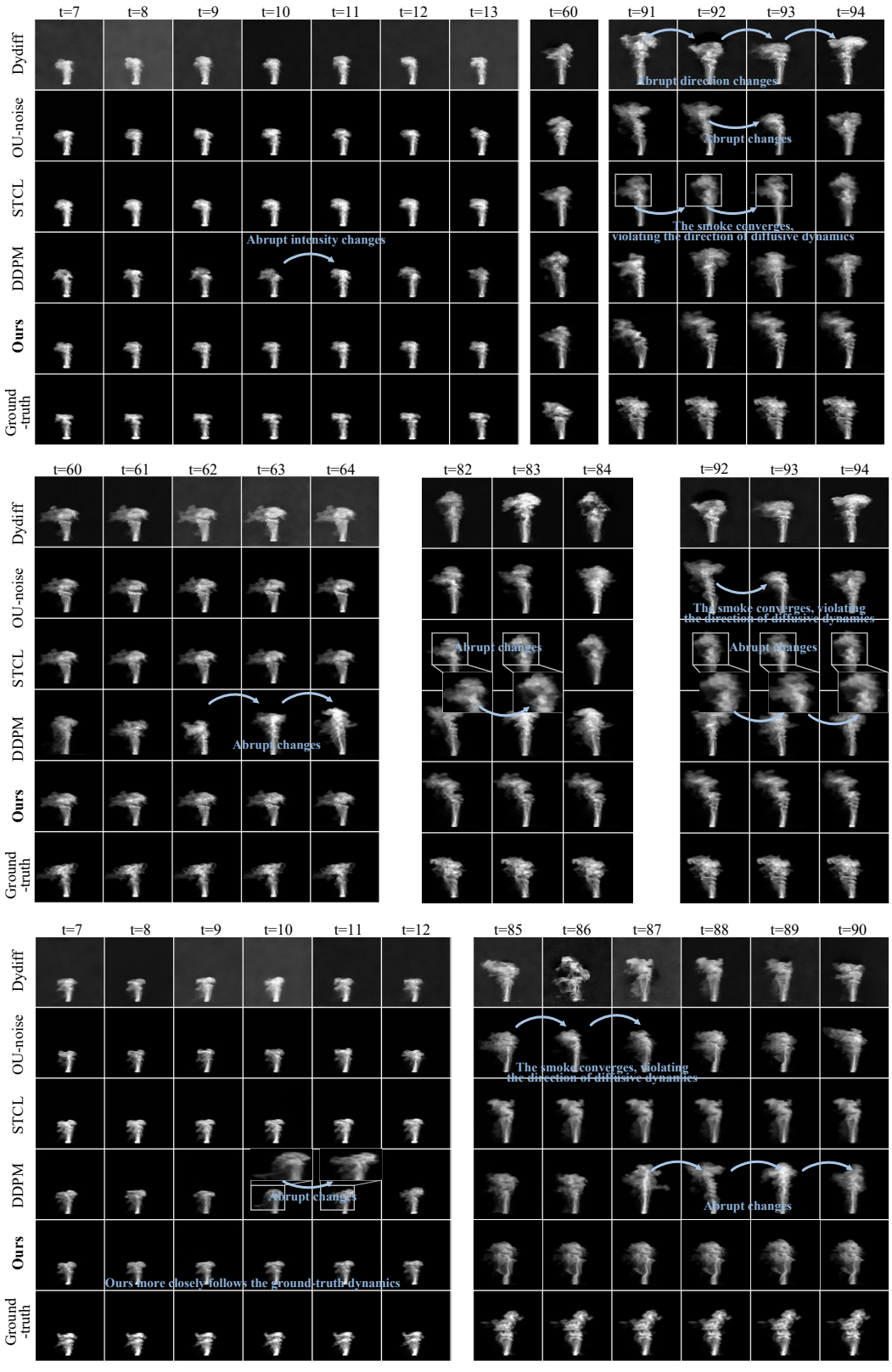


Figure 11: **Visualization** in Scalarflow shows that our predictions better follow the ground-truth trend, preserving continuous and consistent smoke dispersion.