Fitting Image Diffusion Models on Video Datasets

Juhun Lee
Dept. of Artificial Intelligence
Sungkyungwan University
josejhlee@g.skku.edu

Simon S. Woo*
Dept. of Artificial Intelligence
Sungkyungwan University
swoo@g.skku.edu

Abstract

Image diffusion models are trained on independently sampled static images. While this is the bedrock task protocol in generative modeling, capturing the temporal world through the lens of static snapshots is information-deficient by design. This limitation leads to slower convergence, limited distributional coverage, and reduced generalization. In this work, we propose a simple and effective training strategy that leverages the temporal inductive bias present in continuous video frames to improve diffusion training. Notably, the proposed method requires no architectural modification and can be seamlessly integrated into standard diffusion training pipelines. We evaluate our method on the HandCo dataset, where hand-object interactions exhibit dense temporal coherence and subtle variations in finger articulation often result in semantically distinct motions. Empirically, our method accelerates convergence by over 2x and achieves lower FID on both training and validation distributions. It also improves generative diversity by encouraging the model to capture meaningful temporal variations. We further provide an optimization analysis showing that our regularization reduces the gradient variance, which contributes to faster convergence.

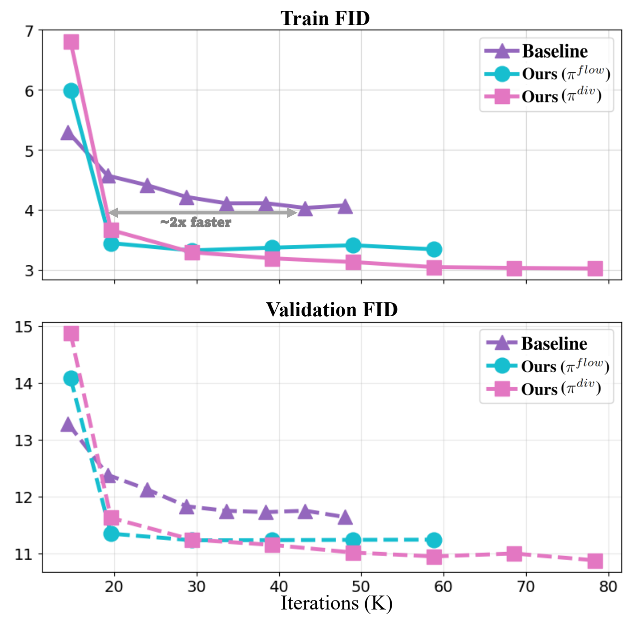

Figure 1. FID scores over training iterations for the baseline and our method variants. Our approach converges faster and consistently achieves lower train and validation FID, demonstrating improved training efficiency and generalization performance.

1. Introduction

Understanding and advancing the capabilities of diffusion models hinges, in large part, on how effectively they are fitted to data. To date, the community's dominant focus for such study has been static image datasets, and success on this modality has become a cornerstone benchmark for theoretical and empirical progress in visual generative modeling.

In this work we deliberately depart from that convention and examine an alternative training configuration: training image diffusion models on temporal (video) datasets. There are two principal motivations for embracing this configuration.

Firstly, when every frame in a video is itself a valid and semantically meaningful image, the temporal stream forms a semantic traversal through the underlying data manifold. A hand-gesture dataset, for example, yields a sequence in which even the slightest articulation encodes a meaningful shift in semantics. By treating each frame as a legitimate training instance while simultaneously contextualizing it with its temporal neighbors, we can expose the model to a structured prior that is entirely absent when frames are shuffled into an i.i.d. pool. In effect, static snapshots of the world distribution are now embedded in a coherent local neighbourhood that supplies the model with auxiliary cues

* Corresponding author

about continuity, shape, and appearance variation.

Another motivation is rooted in training stability and convergence behavior. Diffusion models are notorious for slow convergence, and under data-scarce regimes, they often over-fit to a subset of "easy" modes. This pathology arises from an intrinsic coupling between the denoising objective and the accumulation of sampling error: a perfect denoiser would, in principle, memorise the training data, whereas, in practice, the small errors incurred at each reverse diffusion process step both "bless" the model with the ability to generate novel samples and "doom" it to preferentially denoise towards most readily reconstructible samples. Prior successes on data augmentation in GAN training (e.g. ADA, Diff-Aug, negative augmentation) suggest data augmentation can regularise diffusion training, temper over-fitting, and foster richer representation learning.

For these reasons, our work quantifies the influence of the data geometry presented during training and treats it as a controllable variable for both optimisation stability and convergence optimality. Concretely, we introduce a proximity-weighted Laplacian smoothing term that penalises discrepancies between predictions of temporally adjacent noisy inputs. To even our surprise, this single regularization yields markedly faster convergence (2x) and superior generative quality (a whole -1.0 train FID) on the training set, while simultaneously improving generalisation (-0.87 val. FID) beyond mere memorisation of seen data. Furthermore, our optimization analysis reveals that the regulariser reduces gradient variance by aligning local Jacobians, thereby explaining its stabilising effect. We hope that this study inspires further exploration of the data set structure as a lever to shape the training dynamics of diffusion models.

2. Related Works

Denoising diffusion probabilistic models (DDPMs) [4, 5, 9] have emerged as a powerful class of generative models, achieving state-of-the-art performance in image synthesis, editing, and conditional generation tasks. These models learn to reverse a gradual noising process through iterative denoising steps, guided by a neural network trained to predict added noise [5, 9]. Variants have explored architectural improvements [14] and applications to various modalities such as text [21], video [11, 15, 18]. Complementary to these advances, recent work has sought to accelerate the notoriously long training cycles of diffusion models: the Min-SNR weighting strategy [8] adaptively prioritizes low-signal-to-noise training examples, while FasterDiT [20] introduces architectural and algorithmic refinements that markedly shorten convergence time.

Recent studies have begun to examine the generalization ability of diffusion models [12]. Some approaches introduce regularization schemes [7, 19] to improve robustness to unseen samples or encourage smoother latent spaces. Others analyze mode coverage, memorization, and diversity [1, 13], highlighting that diffusion models may still suffer from mode collapse or overfitting under limited or biased training data. Recent representation-learning regularizers provide orthogonal benefits: REPresentation Alignment (REPA) aligns noisy-frame features with pretrained encoder embeddings to speed up diffusion-transformer training [17], while Dispersive Loss repels hidden activations in a positive-free contrastive style, boosting ImageNet FID without extra data [17]. Both act purely in latent space, complementing our temporal-proximity regularizer. However, explicit use of temporal information to improve generalization has received limited attention, especially in the context of image diffusion model training. Also, due to the "memorizing" nature of the diffusion objective, regularizing the model to restrain itself from memorizing has direct implications for privacy preservation in diffusion models [6, 10].

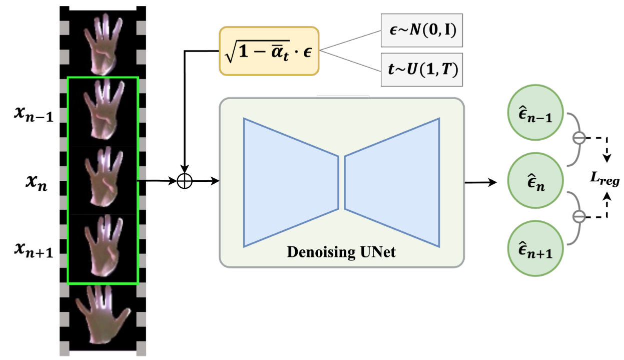

Figure 2. Overview of our regularization term. Three adjacent video frames are corrupted with shared noise and timestep. A temporal consistency regularization loss is applied to enforce consistency across predicted noise outputs.

3. Method

Given a collection of static images $\mathcal{D}_{\text{img}} = \{x_i\}_{i=1}^N \subset \mathbb{R}^d$, the empirical measure is the Dirac mixture:

$$\mu_{\text{img}} =: \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}, \tag{1}$$

where δ_{x_i} is the Dirac delta distribution centered at sample x_i. Consequently, it provides the minimizer with no explicit notion of neighbourhood or smooth variation across data. Under an over-parameterised network, minimizing i.i.d samples with the diffusion objective encourages memorisation: the model interpolates training samples without regard for their relative arrangement in the ambient space $\mathbb{R}^d$, often harming generalisation to unseen data.

Consider a video dataset expressed as continuous trajectories $\{\gamma_v : [0, T_v] \to \mathbb{R}^d\}_{v=1}^V$. Integrating Dirac deltas along each path yields the trajectory–supported empirical measure:

$$\mu_{\text{vid}} =: \frac{1}{\sum_{v=1}^{V} T_v} \sum_{v=1}^{V} \int_0^{T_v} \delta_{x-\gamma_v(t)}, dt, \tag{2}$$

which carries the natural concept of proximity—adjacent frames lie infinitesimally apart along the same curve. Importantly, note that even if one were to train with μ_{vid}, the exploitation of such a temporal structure is not explicitly encouraged. The interplay between data shuffling in mini-batches and overparameterized networks can still admit memorising solutions. Explicit regularisation is therefore necessary to translate the inductive bias in (2) into well-generalizing smoother hypotheses.

3.1. Problem Definition

Let $(x_1, x_2, \ldots, x_N) \in \mathbb{R}^{N\times C\times H\times W}$ denote a sequence of N consecutive video frames, where each frame x_i is a clean image with C channels and a spatial resolution of $H \times W$. For each target frame x_n, we define a temporal window (x_{n-1}, x_n, x_{n+1}) consisting of the frame itself and its adjacent frames in time. The goal of this study is to train a model that leverages this local temporal context to generate temporally consistent predictions for the target frame x_n.

We adopt the standard noise prediction formulation used in denoising diffusion probabilistic models (DDPMs) [9]. During training, we perform the forward diffusion process on clean images $x \sim p_{\text{data}}(x)$ using a randomly sampled timestep $t \sim \text{Unif}(1, T)$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$:

$$\tilde{x}_t = \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{3}$$

where $\bar{\alpha}_t$ denotes the cumulative product of noise scaling coefficients. The denoising model ϵ_θ is trained to predict the added noise using a mean squared error loss [9]:

$$\mathcal{L}_{\text{mse}} = \mathbb{E}_{x,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\tilde{x}_t, t)\|_2^2\right]. \tag{4}$$

However, this basic objective treats each frame independently and fails to directly capture local frame density. To overcome this limitation, we propose a training strategy that exploits temporal information within the triplet frame window to enforce consistency at the noise level.

3.2. Shared Epsilon-Noise Injection

The vanilla diffusion objective perturbs every training frame independently, drawing both the diffusion timestep t and the Gaussian noise ϵ afresh for each image. That independence makes adjacent frames unrelatable: even if two inputs depict almost identical content, the network observes them through distinct noise realisations at different noise levels. To this end, before any consistency term can be imposed, we align the stochastic conditions across a local window of K consecutive frames. Concretely we sample

$$\tau \sim \text{U}(0, T), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

once per window and inject the same pair $(\tau, \boldsymbol{\varepsilon})$ into every frame x_i in that window:

$$\tilde{x}^i \;=\; \sqrt{\bar{\alpha}_\tau}\, x^i \;+\; \sqrt{1-\bar{\alpha}_\tau}\, \boldsymbol{\varepsilon}, \qquad i = 1, \ldots, K. \tag{6}$$

This strategic adjustment leaves the architecture untouched and introduces no extra parameters, yet it synchronises the noise signature and its magnitude across neighbouring frames. Under this aligned structure, we gain access to apply intra-sample objectives.

3.3. Temporal Consistency Regularization

Now, we wish to explicitly inject the geometric structure of the video measure μ_{vid} into the learning objective, which should carry an intrinsic notion of proximity along each trajectory. We introduce a generic, symmetric, positive weighting function $w_{i,j} = \varphi\left(\pi_{i,j}\right)$ that maps any scalar proximity measure $\pi_{i,j} \in \mathbb{R}_{\geq 0}$ between frame x^i and x^j into a Laplacian weight. Throughout the paper φ is taken to be a monotone decay function ($\varphi' < 0$), so that two frames judged closer receive a larger coupling.

Section 3.2 introduced the shared (τ, ε) injection so that each frame in a window is observed under identical corruption. This design equalises scale and stochasticity across the window, letting any inter-frame operation possible—an essential pre-condition for our regularization term. Namely, given the aligned noisy inputs $\tilde{x}^i$ and their noise predictions $\hat{\varepsilon}^i_\theta = \varepsilon_\theta(\tilde{x}^i, \tau)$, we penalize disparities between adjacent predictions:

$$\mathcal{L}_{\text{reg}} = \sum_{i=0}^{N-2} w_{i,j} \left\|\hat{\epsilon}^i_\theta - \hat{\epsilon}^j_\theta\right\|_2^2, \tag{7}$$

with $j = i + 1$. The weight $w_{i,i+1}$ acts as a discrete, flow-aware Laplacian, turning $\mathcal{L}_{\text{reg}}$ into a proximity-weighted Dirichlet energy that directly encodes the neighbourhood structure implicit in μ_{vid}. Finally, our training loss is the composite objective

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mse}} \;+\; \lambda\, \mathcal{L}_{\text{reg}}, \tag{8}$$

Next, we concretize two choices for $\pi_{i,i+1}$

3.4. Optical-Flow Proximity

Following classical motion analysis, we set $\pi_{i,j}$ to the mean-squared optical-flow magnitude between consecutive frames:

$$\pi^{flow}_{i,j} = \frac{1}{HW} \sum_p \left\|F_{i\to j}(p)\right\|_2^2, \tag{9}$$

where F is any off-the-shelf flow estimator and p indexes pixels. Choosing $\varphi(\pi) = \left(\pi + \delta\right)^{-1}$ with a small $\delta > 0$ produces the weight of Eq. (7), which serves as a data-driven proxy for geodesic distance along the trajectory manifold μ_{vid}.

3.5. Trajectory Divergence Proximity

We consider an estimator-free alternative that relies solely on the model's own forward process. The presentation of such is to both highlight a computationally friendly approach to our regularization and convince the flexibility of the design of the proximity.

Let $d^{i,j} = \frac{1}{CHW}\left\|x_t^i - x_t^j\right\|_2^2$ be the normalised squared distance between two noisy frames obtained at the same diffusion step t. We approximate its time derivative by a symmetric finite-difference:

$$\pi_{i,j}^{div} =: \dot{d}^{i,j} \;\approx\; \frac{d_{t+\Delta t}^{i,j} - d_{t-\Delta t}^{i,j}}{2\,\Delta t}, \tag{10}$$

with a small integer step Δt. Operationally, the four noisy realisations $x_{t\pm\Delta t}^{i,j}$ are obtained in a single forward process call, making the computation lightweight. Intuitively, $|\dot{d}|$ measures the rate in which the two trajectories diverge (or converge) in latent space: frames whose distance changes slowly are deemed proximal. As for the weighting function, we set $w_{i,j} = \frac{1}{\varepsilon + |\pi_{i,j}|^{\frac{1}{2}}}$.

4. Optimization Analysis

For simplicity of our following derivations, we introduce additional notations. Model each training sample as three consecutive frames x_0, x_1, x_2 and denote their UNet noise–predictions by the d-dimensional vectors $f_i := \epsilon_\theta(x_i) \in \mathbb{R}^d$ ($d = C \times H \times W$). For any parameter coordinate θ we write the corresponding Jacobian rows $J_i := \partial_\theta \epsilon_\theta(x_i) \in \mathbb{R}^d$. Edge-wise quantities are $s_{ij} := f_i - f_j$ (aka. output mismatch) and $D_{ij} := J_i - J_j$. Now, our temporal regularizer is a discrete Dirichlet energy on the output matching

$$E_S(\theta) \;:=\; \frac{1}{2} \sum_{(i,j)\in E} w_{ij}\,\|s_{ij}\|_2^2, \qquad \mathcal{L}_{\text{reg}}(\theta) = E_S(\theta), \tag{11}$$

Similarly, we can define the Dirichlet energy on the Jacobian matching level

$$E_G(\theta) \;:=\; \frac{1}{2} \sum_{(i,j)\in E} w_{ij}\,\|D_{ij}\|_F^2, \tag{12}$$

which measures the roughness of the parameter-space sensitivity field over the local input graph. Let ℓ_i be the per-sample denoising loss at x_i. A first-order expansion gives the exact decomposition

$$\nabla_\theta \ell_i - \nabla_\theta \ell_j \;=\; J_i^\top s_{ij} \;+\; D_{ij}^\top f_j. \tag{13}$$

Assumption (Uniform bounds on the local input set). Let $\mathcal{N} := \{x_i\}_{i=1}^N$ denote the neighborhood/window used to form E. There exist finite constants $G, F > 0$ such that

$$\sup_{x\in\mathcal{N}} \left\|\tfrac{\partial f(x,\theta)}{\partial\theta}\right\|_2 \le G \quad \text{and} \quad \sup_{x\in\mathcal{N}} \|f(x,\theta)\|_2 \le F.$$

With these pre-set bounds, we can contain the difference norm of adjacent gradients in Eq. (13) as

$$\|\nabla_\theta \ell_i - \nabla_\theta \ell_j\|_2^2 \;\le\; 2G^2\,\|s_{ij}\|_2^2 \;+\; 2F^2\,\|D_{ij}\|_2^2. \tag{14}$$

In other words, jointly matching outputs and Jacobians suppresses per-sample gradient disagreements. While the disagreement $\|s_{ij}\|$ is explicitly minimized by enforcing regularization, the first-order approximation of the updated Jacobians is

$$J_i^+ \approx J_i - \eta\, H_j\, \nabla_\theta \mathcal{L}_{\text{reg}} \tag{15}$$

Yielding the difference of two adjacent Jacobians J_i and J_j, we have

$$D_{ij}^+ \;\approx\; D_{ij} - \eta \underbrace{(H_i - H_j)}_{\text{frames' curvature gap}} \nabla_\theta \mathcal{L}_{\text{reg}} \tag{16}$$

The update subtracts a component of D_{ij} driven by its correlated portion with the current discrepancy s_{lm}. As those discrepancy directions vary across the updates, the regularization continuously remove this very correlation and $||D_{ij}||$ shrinks.

Now let L_w be the (unnormalized) graph Laplacian on the local graph $(\{1, \ldots, N\}, E)$ with weights w_{ij}, and

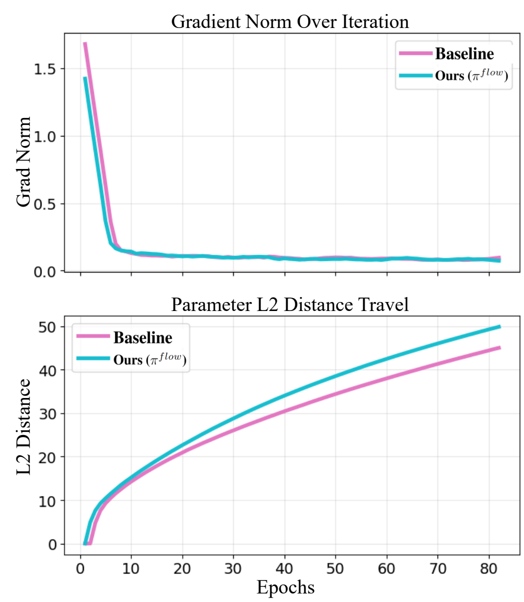

Figure 3. Gradient norm and L^2 parameter distance, compared across baseline and our (π^{flow} variant) model.

$\lambda_2(L_w) > 0$ its algebraic connectivity (aka. the second smallest eigenvalue). For node-attached vectors $u_i \in \mathbb{R}^P$ with mean $\bar{u} = \frac{1}{N}\sum_i u_i$, the discrete Poincaré inequality states

$$\frac{1}{N}\sum_{i=1}^{N}\|u_i - \bar{u}\|_2^2 \;\le\; \frac{1}{N\,\lambda_2(L_w)} \sum_{(i,j)\in E} w_{ij}\,\|u_i - u_j\|_2^2. \tag{17}$$

Choosing $u_i = \nabla_\theta \ell_i$ and applying (14) yields the explicit variance bound

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N}\|u_i - \bar{u}\|_2^2}_{\mathrm{Var}(\nabla_\theta \ell)} \;\le\; \frac{4}{N\,\lambda_2(L_w)}\Big(G^2\,E_S(\theta) + F^2\,E_G(\theta)\Big), \tag{18}$$

where $\overline{u} := \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \ell_i$ is the batch-mean (global) gradient. Equation (18) formalizes a compact message: local regularization that matches outputs (E_S) and enforces Jacobian similarity (E_G) over adjacent inputs directly controls the global dispersion of per-sample gradients around their batch mean. Since SGD uses the batch mean as its update direction, reducing $\mathrm{Var}(\nabla_\theta \ell)$ improves the signal-to-noise ratio of the update, producing straighter optimization trajectories and faster, more stable convergence. Empirically, we calculate the average gradient norm for 50 epochs for both the baseline and our model. In tandem, we measure the l2 parameter distance between the updating parameter and its initialization, shown in Fig. 3. Notice that with a comparable gradient norm budget, the model parameter travels farther.

5. Experiments

Experimental Setup. We evaluate our method using the HandCo dataset [2, 3], which consists of hand gesture video sequences captured in controlled environments, as it contains abundant temporal variation due to frequent finger and hand articulation over time, making it well-suited for evaluating the impact of temporal supervision. Moreover, the nature of hand motion in video often leads to transient occlusions or anatomically ambiguous frames—such as missing or overlapping fingers—which can challenge static supervision methods and highlight the benefits of temporally informed training. To solely focus on the gesture itself, we mask the background in all frames. All models are trained with the epsilon-prediction objective (DDPM). During sampling, we iterative 100 sampling steps with DDIM [16]. We use a U-Net convolution-attention hybrid backbone for the denoising network across all experiments.

Training Variants. To assess the benefit of explicit regularisation, we compare across varied configurations. All settings use the same video dataset.

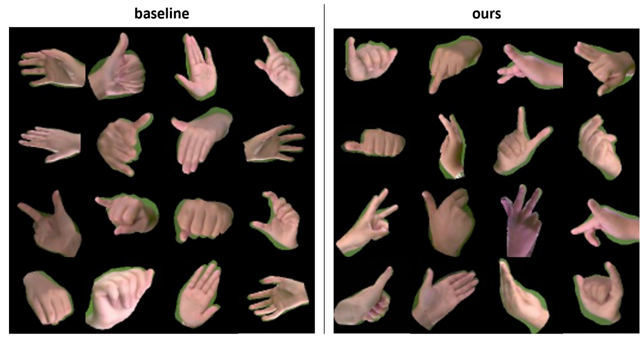

Figure 4. Qualitative comparison of generated images. Left: samples from the baseline model. Right: samples generated by our (π^{flow} variant) method. Our model produces more diverse outputs.

- **Baseline**: frames are sampled independently across the entire dataset, obliterating temporal structure.
- **Sequence-preserving**: sampling is i.i.d. at the video level, while frame order within each clip is retained, exposing short trajectories but without coupling across them. This is to encourage noisy variation canceling across similar frames in the batch.
- **Adjacent consistency**: We ablate the weighting in our regularization.
- **Dispersive Loss**: We apply a representation learning-centric regularization loss, which can be understood as a contrastive loss without penalizing positive terms. For more details, please refer to their work [17].
- **Ours** (π^{flow}): Our regularization with optical flow-based proximity.
- **Ours** (π^{div}): Our regularization with trajectory divergence proximity with $\Delta t = 50$.

For all baselines, we train a standard DDPM using static single-frame supervision ($seq\_len = 1$) with a batch size of 256 for 500 epochs. The learning rate is set to 0.0001, and exponential moving average (EMA) with a decay of 0.9995 is applied. For the dispersive loss, we set the temperature at 0.5, with $\lambda = 0.005$ For our methods, we adopt the same architecture and diffusion configuration but extend the input sequence length to three consecutive frames ($seq\_len = 3$). We set $\lambda = 0.1$. The batch size is set to 128 and training is conducted for 300 epochs. Both models are trained with 25,000 samples randomly selected from the HandCo training set. During evaluation, we generate 50,000 images and report results based on FID and qualitative sample diversity. All experiments are conducted using 4 NVIDIA A5000 GPUs.

5.1. Experimental Results

Tab. 1 presents the FID scores on both the training and validation sets for the baseline and our proposed method.

FID-25k	Train	Val
Baseline	4.02	11.74
Seq. Preserving	3.98	11.70
Adj. Consistency	4.59	11.67
Dispersive Loss	4.68	12.25
Ours (π^{flow})	<u>3.31</u>	<u>11.23</u>
Ours (π^{div})	**3.02**	**10.87**

Table 1. FID scores for baseline and our method. Lower is better. We report performance of models (at their best) at 450-500 training epochs, while our method achieves its best performance at early as 150 epochs.

The baseline reaches its best performance at 450 training epochs, whereas our method achieves superior FID scores after only 150 epochs. This highlights that our temporally grounded training strategy not only enhances generative quality and generalization but also accelerates convergence. Furthermore, as illustrated in Fig. 3, our method achieves comparable or better validation FID scores approximately 2.5× faster than the baseline, clearly demonstrating its training efficiency and effectiveness.

Fig. 4 presents a qualitative comparison between the baseline model and our proposed method. While the baseline model tends to produce visually similar or repetitive outputs, our model generates a broader range of appearances with greater variation in pose, shape, and orientation. This suggests that the proposed temporal supervision strategy encourages the model to fairly treat samples of varying modes equally.

6. Discussion

6.1. Ablation Studies

We study the convergence behavior by ablating and highlighting exclusive components. Namely, we notice that the preservation of the sequence during parameter update incrementally helps on convergence. As hypothesized, we believe this advantage arises from the averaging of noisy gradients per sequence, resulting in higher fidelity sequence-wise gradients. However, even with such a structured sampling, we see that it is not sufficient for any significant improvement over the baseline. For the case of adjacent consistency, where we ablate the proximity weighting, the model suffers noticeable performance degradation: compared to ours (π^{div}), the validation FID increases from 10.87 to 11.67, and the training FID from 3.02 to 4.59. This confirms that flow-guided weighting plays a crucial role in offering the right Laplacian weighting matrix. Interestingly, Dispersive Loss doesn't have a clear positive contribution, both in train and validation FID. Lastly, we verify that while both designs achieve superior performances, π^{div}) show stronger model fitting (Train FID) and generalization (Validation FID).

6.2. Limitations

Our formulation presumes that consecutive frames trace smooth trajectories on the data manifold (Sec. 3.3). In settings with abrupt shot changes, strong camera shake, or heavy motion blur, the optical-flow magnitudes or the empirical trajectory divergence used to weight the Laplacian can become unreliable, leading either to vanishing weights or—in the worst case—to spurious penalties that harm convergence. Robust flow estimation—or alternative self-supervised proximity signals—will be required before the method can be safely deployed on unconstrained "in-the-wild" video. In regards to the optimization analysis, the variance-reduction argument assumes idealized training conditions (e.g. bounded Hessians and Polyak-Lojasiewicz (PL) Condition). While the empirical trends support the analysis, stronger guarantees remain open.

7. Conclusion

We introduced a proximity-based regulariser for diffusion models that ties together the denoising predictions of neighbouring frames under shared noise by using temporally structured video datasets. The weight ω is deliberately agnostic to any one metric: it can be driven by optical-flow magnitude or, as we show, by a new latent trajectory divergence that looks at how inter-frame distance changes along the diffusion path. When applied to hand-gesture data, this "frame-bridging" constraint forces the generator to respect the subtle pose shifts and finger articulations that separate one gesture phase from the next, improving both visual fidelity and temporal coherence. The approach requires no external trackers, adds negligible cost, and offers a principled template for injecting fine-grained temporal semantics into generative training. Empirically, its effectiveness is pronounced in both train FID and validation FID through better fitting without overfitting.

Acknowledgments

We acknowledge the help and contribution of the following people: Yulim So for the optical flow calculation, method section writing, and diagram drawing, Sehun Chang for plotting the results, Soobin Cha for reviewing the paper, and Jae-Pil Heo for providing the guidance for the project. This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (RS-2022-II221199, RS-2022-II220688, RS-2019-II190421, RS-2023-00230337, RS-2024-00356293, RS-2024-00437849, RS-2021-II212068, RS-2025-02304983, and RS-2025-02263841).

References

[1] Sumukh K Aithal, Pratyush Maini, Zachary Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in Neural Information Processing Systems*, 37:134614–134644, 2024. 2

[2] Jimei Yang Bryan Russell Max Argus Christian Zimmermann, Duygu Ceylan and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 5

[3] Max Argus Christian Zimmermann and Thomas Brox. Contrastive representation learning for hand shape estimation. In *arxive*, 2021. 5

[4] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 2

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[6] Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023. 2

[7] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7558, 2024. 2

[8] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7441–7451, 2023. 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[10] Seunghoo Hong, Juhun Lee, and Simon S Woo. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21143–21151, 2024. 2

[11] Emiel Hoogeboom, Jonathan Ho, Mohammad Norouzi, and Tim Salimans. Video diffusion models. In *CVPR*, 2023. 2

[12] Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. *arXiv preprint arXiv:2310.02557*, 2023. 2

[13] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 2

[14] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[15] Uriel Singer, Adam Polyak, Thomas Hayes, and et al. Mcvd: Masked conditional video diffusion. In *NeurIPS*, 2022. 2

[16] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[17] Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025. 2, 5

[18] Zirui Wang, Ziyang Ma, Dongxu Li, and et al. Video fusion diffusion for video generation. In *ICLR*, 2024. 2

[19] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 2

[20] Jingfeng Yao, Cheng Wang, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *Advances in Neural Information Processing Systems*, 37:56166–56189, 2024. 2

[21] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2