EUCLIDEAN DISTANCE-BASED FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow Matching has recently become popular as a technique for generative modeling with applications in many domains. In the standard Flow Matching formulation, we interpret the vector field as the solution of a least-squares regression problem, which is known to be sensitive to contamination. Therefore we leverage the viewpoint of robust statistics and replace the squared Euclidean distance with the Euclidean distance as a regression objective. This seemingly minor change strongly affects the properties of Flow Matching. We provide a probabilistic interpretation of the new regression target as a conditional geometric median and use this characterization to study distributional robustness of the Euclidean Distance-based Flow Matching. We analyze the mathematical properties of this changed objective and the effect of a contaminated target distribution on the involved conditional velocity distributions. We discuss the challenge of directly characterizing the distribution that our method induces. In specific contamination settings we show theoretically that our method yields an improvement for severe contamination. Finally, we present numerical experiments conducted on the CIFAR-10 image dataset where a proportion of the samples has been contaminated. Our Euclidean distance-based Flow Matching shows a clear advantage in scenarios of severe contamination of the target distribution.

1 INTRODUCTION

Flow Matching (FM) (Lipman et al., 2023; 2024; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023) is a framework for training a flow model (Chen et al., 2018; Grathwohl et al., 2019) to generate samples from a target distribution that is available only through data. Starting from a known source distribution from which we can sample, FM learns a time-dependent vector field that drives the generation process. New samples are generated by drawing from the source distribution and applying the flow induced by this learned time-dependent vector field. It is well known that under suitable regularity assumptions, the (population) optimal vector field for the standard formulation generates the prescribed target distribution exactly.

Real-life datasets often do not only contain samples from the assumed true distribution, but they contain contaminated samples as well. A classical example is the Belgium telephone calls dataset (Rousseeuw & Leroy, 1987) that includes several wrongly recorded data points. Therefore we model the observed distribution as a mixture model of the clean and a contaminating distribution, rather than only the target distribution. Consequently, robustness is the desirable property that the learned distribution stays close to the clean distribution, even in the presence of contamination without manually filtering the samples.

Flow Matching is a powerful technique to produce samples from the target, but suffers from replication of the exact target including the contamination. We present a remedy for this by a simple modification of the Flow Matching objective. Our approach replaces the squared Euclidean distance loss in the standard Flow Matching objective with the Euclidean distance loss. The latter penalizes large regression residuals linearly rather than quadratically, motivating its use to reduce sensitivity to contamination. At the population level, the modified field is a geometric median of the conditional velocity distribution, whereas the standard squared-loss field is its mean. We demonstrate that this change also improves distributional robustness, yielding a generated distribution closer to the clean target under contamination. Clearly the necessary trade-off that we accept is that we lose the general guarantee that the population-optimal field reproduces the prescribed target distribution, even when the data are uncontaminated.

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.

To be more precise, we characterize the population-optimal Euclidean distance Flow Matching vector field as a conditional geometric median. On clean data, we derive bounds comparing this field with the standard squared-loss field. We then analyze how contamination affects the regression fields and prove that, in specific contamination settings, our adaptation yields an improvement in the Wasserstein distance. Finally, we evaluate both objectives on contaminated CIFAR-10 data, comparing their generation quality across different contamination models and proportions.

1.1 Related Work

Several works focus on improving Flow Matching by modifying the coupling between source and target samples, for instance Pooladian et al. (2023) and Tong et al. (2024a). Repeated rectification also straightens trajectories (Liu et al., 2023). These approaches modify the coupling or transport path while retaining the standard squared regression objective. Song et al. (2025) consider a related alternative objective in visuomotor learning. In their main method, they reformulate velocity prediction as terminal-sample prediction and train the resulting model using a componentwise absolute-deviation loss together with a two-step sampling procedure, although they also test their loss with conventional flow integration. They empirically demonstrate faster convergence on visuomotor-learning benchmarks. We use Euclidean-distance based loss, which is invariant under orthogonal changes of coordinates and yields a conditional geometric median rather than coordinatewise conditional medians. Moreover, our focus is distributional robustness.

Robustness was recently studied in the closely related setting of diffusion models and score matching. Khrapov et al. (2024) replace the standard squared denoising objective with a scheduled pseudo-Huber loss with a time-varying threshold. They empirically demonstrate improved resistance to contaminated training data in both image and speech generation tasks. Similarly Li & Dong (2026) study replacing the mean squared error denoising loss with transformations induced by f-divergences. Their framework includes bounded-influence choices that downweight samples with large residuals and empirically improve generation under contamination. In the distinct setting of parametric score matching, Schwank et al. (2025) construct a robust estimator through geometric median of means aggregation and establish theoretical guarantees with particular attention to exponential-family models. Our approach differs in that the geometric median is not introduced as an external aggregation procedure but instead arises directly as the population-optimal regression target of our modified objective.

2 Technical Background

Let us briefly recall technical foundations of Flow Matching. Denote the source distribution as P_0 and the target distribution as P_1. Throughout we assume the source to be Gaussian $P_0 = \mathcal{N}(0, \mathrm{I}_n)$. We follow the interpolation path $(P_t)_{t\in[0,1]}$ which is the linear path interpolating independent random vectors (Peyré, 2026):

$$P_t = \mathrm{Law}(X_t), \quad \text{where} \quad (X_0, X_1) \sim P_0 \otimes P_1\,, X_t = (1-t)X_0 + tX_1.$$

Throughout we assume that all endpoint distributions have finite second moments. The goal is to find a time dependent vector field $v : [0,1] \times \mathbb{R}^n \to \mathbb{R}^n$, $(t,x) \mapsto v_t(x)$ such that the pair (P_t, v_t) satisfies the continuity equation (technically understood in the weak sense)

$$\partial_t P_t + \mathrm{div}(v_t P_t) = 0.$$

The flow $\phi : [0,1] \times \mathbb{R}^n \to \mathbb{R}^n$ induced by $v_t(x)$ is defined as the solution to the following ordinary differential equation (ODE)

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(x) = v_t(\phi_t(x)), \quad \text{with } \phi_0(x) = x.$$

If $v_t(x)$ is sufficiently regular it follows from Villani (2009) that its induced flow generates the previously defined probability path P_t in the sense that $(\phi_t)_{\#}P_0 = P_t$. In particular, it follows that $\phi_1(X_0) \sim P_1$. Consequently such a vector field enables generating samples from the target distribution, by sampling from the source and integrating the ODE which is done numerically. In standard Flow Matching the population objective is

$$\min_{(v_t)_{t\in[0,1]}} \int_0^1 \int_{\mathbb{R}^n\times\mathbb{R}^n} \|v_t((1-t)x_0 + tx_1) - (x_1 - x_0)\|^2 \mathrm{d}P_0(x_0)\mathrm{d}P_1(x_1)\mathrm{d}t. \tag{1}$$

Throughout the paper, we denote by $v_t^2(x)$ a measurable population minimizer of Equation 1. By the optimal regression property under a squared objective, we obtain that the population-level minimizer can be written as a conditional expectation (Peyré, 2026; Albergo et al., 2025):

$$v_t^2(x) = \mathbb{E}\left[X_1 - X_0 | X_t = x\right]. \tag{2}$$

Then $v_t^2(x)$ and P_t solve the continuity equation in the weak sense. In practice we work with Equation 1 which is tractable via empirical risk minimization. Using a simple source distribution we can generate samples from P_0 and the target distribution is available through data samples.

2.1 Euclidean distance regression objective

We aim to study the effects of replacing the squared Euclidean norm in the regression problem $\|\cdot\|^2$ with the norm $\|\cdot\|$. We investigate whether the reduced sensitivity of this regression target to contamination can yield a terminal distribution closer to the clean target when the observed target distribution is contaminated. Lipman et al. (2024) show that the optimal vector field after replacing the loss $\|\cdot\|^2$ in Equation 1 by a Bregman divergence, applied with the target velocity as the first argument and the prediction as the second, preserves the probability path. Our proposed loss is not covered by this and we have no guarantee that our method generates the probability path P_t. In particular, we cannot guarantee that the learned target coincides with the observed target. We highlight that this is an important property of our method, since we do not want to replicate contaminated distributions.

Removing the square in Equation 1 our regression objective becomes

$$\min_{(v_t)_{t\in[0,1]}} \int_0^1 \int_{\mathbb{R}^n \times \mathbb{R}^n} \|v_t((1-t)x_0 + tx_1) - (x_1 - x_0)\| \mathrm{d}P_0(x_0) \mathrm{d}P_1(x_1) \mathrm{d}t. \tag{3}$$

We denote by $v_t^1(x)$ a measurable population minimizer of Equation 3. The integral can be seen as a mixed norm and consequently we will also call the method $L_{2,1}$ Flow Matching.

2.2 Discussion of mathematical details

All of our proofs can be found in the supplementary material A. Since conditional distributions are unique only P_t-almost everywhere, characterizations of vector fields as quantities of conditional distributions, as in Equation 2, only hold for Lebesgue-almost every t and P_t-almost every x. The same holds for all pointwise identities and inequalities involving these vector fields. When the conditional geometric median, which is introduced below, is not unique, $v_t^1(x)$ denotes a chosen measurable population minimizer. Note that even when the conditional-mean field admits a continuous representative, the same does not need to hold for the geometric-median field. Although the almost-everywhere characterizations are sufficient for the population regression problems, they do not by themselves identify regular representatives whose associated ODEs are well posed. The existence and uniqueness of induced flows are separate questions that require choosing suitable representatives and imposing additional regularity assumptions. Whenever we analyze an induced distribution, we therefore either construct a particular measurable representative explicitly or state the required existence and regularity assumptions. We make no general claims concerning ODE well-posedness for the $L_{2,1}$ population minimizer.

3 Geometric Median Characterization

In this section we first introduce the geometric median and use it to characterize the population minimizer for the $L_{2,1}$-Flow Matching objective. The geometric median properties used in this section are classical and we include their proofs for convenience. We then apply these properties of the geometric median to help us to compare the resulting vector field with the standard Flow Matching vector field. In one dimension, the median is a measure of location that is more robust to contamination than the mean. The generalization to higher dimensions is not straightforward and multiple generalizations have been suggested (Small, 1990). One possible way to do so is the geometric median (Haldane, 1948; Minsker, 2015). For a probability measure μ on $\mathbb{R}^n$ with finite first moment define its geometric median as

$$\mathrm{Gmed}(\mu) = \arg\min_{y\in\mathbb{R}^n} \int_{\mathbb{R}^n} \|y - x\| \mu(\mathrm{d}x).$$

Its existence is established in Minsker (2015). If the distribution is not concentrated on an affine line, this object is unique (Milasevic & Ducharme, 1987). We will give a probabilistic interpretation of the solution of Equation 3, similar to the mean characterization of squared Flow Matching. For a random variable X we write

$$\mathrm{Gmed}(X) \coloneqq \mathrm{Gmed}(\mathrm{Law}(X))$$

and similarly for a conditional distribution, we write

$$\mathrm{Gmed}(X \mid Z = z) \coloneqq \mathrm{Gmed}(\mathrm{Law}(X \mid Z = z)).$$

3.1 Comparison with standard Flow Matching

Now we want to connect the geometric median to our proposed Flow Matching method and analyze the difference in the resulting vector fields.

Lemma 3.1. *It holds that*

$$v_t^1(x) \in \mathrm{Gmed}(X_1 - X_0 | X_t = x),$$

for almost every t and P_t-almost every x. Moreover, if this conditional distribution is not concentrated on an affine line for almost every t and P_t-almost every x, then the minimizer is unique up to almost everywhere equality.

While this result yields a probabilistic interpretation of the $L_{2,1}$-regression objective, we are unable to give a characterization of the general induced probability path. The proof for standard Flow Matching relies on the tower property of conditional expectations, which to our knowledge does not have a counterpart for the geometric median. Consequently, the standard Flow Matching argument does not show that the $L_{2,1}$-population field generates the prescribed interpolation, and our analysis does not provide a general explicit description.

In the following we analyze the difference between the two vector fields for special scenarios. We first derive that for symmetric distributions, the two vector fields coincide.

Lemma 3.2. *Assume that, for almost every t and P_t-almost every x, the conditional distribution of $X_1 - X_0$ given $X_t = x$ is not concentrated on an affine line and is centrally symmetric around its mean, i.e., for any random variable having this distribution, say Y, we have that $Y - \mathbb{E}[Y]$ matches, in distribution, $-(Y - \mathbb{E}[Y])$. Then we have that*

$$v_t^1(x) = v_t^2(x)$$

for almost every t and P_t-almost every x.

Generally, the target distribution being symmetric around its mean does not imply that the same holds for the conditional velocity distribution of $X_1 - X_0$ given $X_t = x$, even though the Gaussian source is symmetric. As a sufficient condition, if the target distribution is Gaussian, then we can show that the conditional velocity distributions are also Gaussian and thus we obtain the following:

Corollary 3.3. *If the source distribution is a standard Gaussian distribution $X_0 \sim \mathcal{N}(0, \mathrm{I})$ and the target distribution is nondegenerate Gaussian $X_1 \sim \mathcal{N}(\mu, \Sigma)$ where X_0 and X_1 are independent, then*

$$v_t^1(x) = v_t^2(x)$$

for almost every t and P_t-almost every x.

The setting of Corollary 3.3 seems to be quite special however, it is known to be a good proxy in high-dimensional statistics. More generally, the integrated squared difference between the two vector fields is bounded by the optimal squared Flow Matching regression loss.

Theorem 3.4. *In our $L_{2,1}$ Flow Matching setup we have that:*

$$\int_0^1 \int_{\mathbb{R}^n} \|v_t^1(x) - v_t^2(x)\|^2 P_t(\mathrm{d}x)\mathrm{d}t \leq \int_0^1 \mathbb{E}\|(X_1 - X_0) - v_t^2(X_t)\|^2 \mathrm{d}t$$

where the right-hand side is the optimal squared FM regression loss.

Intuitively this shows that the better the velocity can be predicted from interpolation points, the closer the two fields are to each other. This derived velocity field bound is generally not sharp. For instance, for Gaussian targets we have established that both target vector fields are equal, while our bounds do not reflect this. Under additional regularity assumptions, for instance if $v_t^1(x)$ admits a globally Lipschitz representative the above vector field bound could be translated into a Wasserstein bound between the induced distributions. We do not pursue such flow level estimates here.

4 ROBUSTNESS ANALYSIS

In this section, we study how contamination of the target distribution affects the population vector fields of both the squared and the $L_{2,1}$ method. The robustness of the geometric median has been assessed in various works (Small, 1990). It also appears in more recent works, as an aggregate estimator (Minsker, 2015; Minsker & Strawn, 2024). However, in our Flow Matching setting the geometric median is only applied on a conditional distribution at every time and space position to yield a vector field, whose integration in turn yields the generative model. Therefore the robustness properties are highly non-trivial. We separate our analysis into two levels. First, we study how target contamination affects the conditional velocity distributions and the corresponding population vector fields. Second, we investigate whether this robustness can translate into a generated distribution that is closer to the clean target.

We consider a Huber contamination model (Huber, 1964), i.e. we assume that the true target distribution is P and a proportion of ϵ observations are instead replaced with a contaminating distribution Q:

$$P^{\text{cont}} = (1-\epsilon)P + \epsilon Q, \quad 0 \leq \epsilon < 1/2.$$

A crucial property is given by the geometric median's so-called high finite-sample breakdown point. That is the smallest proportion of observations that must be replaced by arbitrary values to make an estimator arbitrarily large or otherwise uninformative. Its finite-sample replacement breakdown point is $B_{\text{GM}} = \lfloor (N+1)/2 \rfloor / N$ (Lopuhaä & Rousseeuw, 1991). This means that if a proportion smaller than B_{GM} of the N observations is replaced, then the displacement of the geometric median is bounded, regardless of the strength of the contamination. In contrast, replacing even a single observation by a point arbitrarily far from the rest of the data can cause an arbitrarily large displacement of the sample mean. This property motivates our use of the geometric median as a velocity estimator, as it remains stable under a substantial proportion of arbitrarily corrupted observations. However, this property does not mean that the geometric median yields better results in all contamination settings. In particular, the mean of a contaminated distribution may be closer to the clean mean than the geometric median.

4.1 ROBUSTNESS PROPERTIES OF THE GEOMETRIC MEDIAN FIELD

We first study how contamination of the target distribution changes the conditional velocity distributions and the corresponding standard and $L_{2,1}$-vector fields. Since population minimizers are defined only up to almost-everywhere equivalence, we fix measurable representatives v_t^2, v_t^1 for the clean target and $v_t^{2,\epsilon}, v_t^{1,\epsilon}$ for the contaminated target. Let

$$B \sim \text{Bernoulli}(\epsilon), X_P \sim P, X_Q \sim Q, X_0 \sim P_0$$

be mutually independent. Let

$$X_1^{\text{cont}} = (1-B)X_P + BX_Q, \quad X_t^{\text{cont}} = (1-t)X_0 + tX_1^{\text{cont}},$$

and

$$X_t^P = (1-t)X_0 + tX_P \quad , X_t^Q = (1-t)X_0 + tX_Q.$$

For t, let γ_t be a measurable version of the posterior contamination probability

$$\gamma_t(x) = \mathbb{P}\left[B = 1 | X_t^{\text{cont}} = x\right].$$

Intuitively, this quantifies the likelihood that after we have sampled t and the corresponding interpolation point, the sampled endpoint was caused by the contaminating distribution. Using the law of total probability we obtain

$$\begin{aligned} &\text{Law}\left(X_1^{\text{cont}} - X_0 \,\middle|\, X_t^{\text{cont}} = x\right) \\ &\quad = (1-\gamma_t(x))\,\text{Law}\left(X_P - X_0 \,\middle|\, X_t^P = x\right) + \gamma_t(x)\,\text{Law}(X_Q - X_0 \,|\, X_t^Q = x). \end{aligned} \tag{4}$$

This is again a mixture distribution, but its weight is the posterior probability $\gamma_t(x)$ rather than ϵ. Moreover, $\gamma_t(x)$ depends on the geometry and magnitude of the contamination. Consequently, moving the contamination farther from the clean data does not uniformly improve the relative robustness of the geometric median. When $\gamma_t(x) \geq 1/2$, the clean conditional component no longer has a strict

majority, so the robustness guarantee below does not apply. Using the linearity of the mean, we can characterize the influence of contamination on the squared vector field

$$\|v_t^{2,\epsilon}(x) - v_t^2(x)\| = \gamma_t(x)\|\mathbb{E}[X_Q - X_0 \,|\, X_t^Q = x] - \mathbb{E}[X_P - X_0 \,|\, X_t^P = x]\|.$$

We notice that in the norm, only the clean conditional mean is independent of the contamination, whereas both $\gamma_t(x)$ and the contaminated conditional mean depend on the contamination. The difference between the conditional means may become arbitrarily large as the contamination is moved farther from the clean data. To fully understand this, one would need to control the posterior contamination probability $\gamma_t(x)$. On the other hand, for $\gamma_t(x) < 1/2$ we have the following bound, adapted from a general result for the geometric median by Minsker (2015).

Lemma 4.1. *For almost every t and P_t^{cont}-almost every x satisfying $\gamma_t(x) < 1/2$, there exists $C_t(x) \geq 0$, depending only on the clean model and not on the contamination, such that*

$$\|v_t^{1,\epsilon}(x) - v_t^1(x)\| \leq \frac{2(1-\gamma_t(x))}{1-2\gamma_t(x)} C_t(x).$$

For $\gamma_t(x) < 1/2$ the factor $\frac{2(1-\gamma_t(x))}{1-2\gamma_t(x)}$ is positive and finite, although it diverges as $\gamma_t(x) \to 1/2$. Generally, those bounds do not directly compare to each other. The squared vector field identity depends on the outlier magnitude and the posterior contamination probability. The $L_{2,1}$-vector field bound depends on a constant that is independent of the contamination and the posterior contamination probability. This suggests that the $L_{2,1}$-vector field can be less distorted by remote contamination whenever $\gamma_t(x) < 1/2$, but it does not imply that it is uniformly closer to the clean field.

Since v_t^2 solves the continuity equation with the clean probability path, we are interested in how much the contaminated $L_{2,1}$-vector field deviates from it. Using the triangle inequality gives

$$\|v_t^2(x) - v_t^{1,\epsilon}(x)\| \leq \|v_t^2(x) - v_t^1(x)\| + \|v_t^1(x) - v_t^{1,\epsilon}(x)\|.$$

This decomposition separates the discrepancy of the mean and the median from the additional distortion caused by contamination. In Section 3.1 we discussed bounds for the first term, while in this section we have established under which conditions we can control the second term. Since this is usually not enough to provide distributional guarantees, we will now present two special settings where we can control the distributional error.

4.2 Wasserstein bounds for special cases

To quantify the discrepancy between generated and clean distributions, we use the 1-Wasserstein distance. For probability measures μ and ν on $\mathbb{R}^n$ with finite first moments, it is defined as

$$W_1(\mu,\nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\| \, \pi(\mathrm{d}x, \mathrm{d}y),$$

where $\Pi(\mu,\nu)$ denotes the set of couplings of μ and ν. Throughout this subsection, since we use linear interpolation with Gaussian source, for an endpoint distribution μ we fix the posterior

$$\pi_{t,x}^{\mu}(\mathrm{d}y) \propto \exp\left(\frac{-\|x - ty\|^2}{2(1-t)^2}\right) \mu(\mathrm{d}y), \quad (t,x) \in [0,1) \times \mathbb{R}^n,$$

where the right-hand side is normalized to unit mass. This choice allows us to impose the geometric median condition for every t, x. Then given a Borel measurable m^{μ} satisfying $m_t^{\mu}(x) \in \mathrm{Gmed}(\pi_{t,x}^{\mu})$ for every $(t,x) \in [0,1) \times \mathbb{R}^n$ the corresponding pointwise $L_{2,1}$ velocity field is

$$v_t^{1,\mu}(x) = \frac{m_t^{\mu}(x) - x}{1-t}. \tag{5}$$

Throughout this subsection, solutions of the associated ODE are understood in the sense that their paths are locally absolutely continuous and satisfy the ODE almost everywhere on $[0,1)$. We first consider a point-mass clean target with point-mass contamination and show that our proposed method yields a smaller Wasserstein distance to the clean target in this setting.

Theorem 4.2. *Let δ_0 be a Dirac distribution in $\mathbb{R}^n$ centered at 0. Let $0 < \epsilon < 1/2$. For $b \in \mathbb{R}^n \setminus \{0\}$, consider the contaminated distribution*

$$P^{\text{cont}} = (1-\epsilon)\delta_0 + \epsilon\delta_b.$$

For every pointwise $L_{2,1}$-velocity field $v^{1,P^{\text{cont}}}$ coming from a Borel geometric-median selection as defined in Equation 5, the associated ODE admits a solution process $(Z_t)_{t\in[0,1)}$ with $Z_0 \sim \mathcal{N}(0, \mathrm{I}_n)$. Every such solution converges almost surely as $t \nearrow 1$ and its terminal law $\rho_1 = \mathrm{Law}(\lim_{t\nearrow 1} Z_t)$ satisfies

$$W_1(\rho_1, \delta_0) < W_1(P^{\text{cont}}, \delta_0).$$

For a point-mass clean target, the preceding theorem establishes improvement for every nonzero displacement of the contaminating point. We now present our main theoretical result, which establishes a corresponding improvement for bounded clean target distributions when the contaminating point is sufficiently far away.

Theorem 4.3. *Let P be a distribution supported in the closed ball $\overline{B}_1(0)$. Let $0 < \epsilon < 1/2$, and fix a unit vector $u \in \mathbb{R}^n$. For $R > 1$, set $b_R = Ru$ and*

$$P_R^{\text{cont}} = (1-\epsilon)P + \epsilon\delta_{b_R}.$$

For each $R > 1$, let $v^{1,R}$ be any pointwise $L_{2,1}$-velocity field associated with P_R^{cont} as defined in Equation 5. Assume that its associated ODE admits a solution process with initial distribution $\mathcal{N}(0, I_n)$. Let $(Z_t^R)_{t\in[0,1)}$ be any such solution process, and write $\rho_t^R = \mathrm{Law}(Z_t^R)$. Then, for all sufficiently large R,

$$\limsup_{t\nearrow 1} W_1(\rho_t^R, P) < W_1(P_R^{\text{cont}}, P).$$

Standard FM aims to reproduce the observed target distribution, which in our setting includes contamination. Our theorems establish that the robustness of the geometric median can translate into a smaller distributional error, in comparison to the contaminated target. For a point-mass target, improvement holds for every nonzero contaminating point. Our main result establishes improvement for bounded target distributions if the contaminating point is sufficiently far away. In this more general setting, the limit superior means that the improvement holds for all times sufficiently close to the final time. This formulation does not require convergence to the final distribution.

5 Experiments

We compare classical Flow Matching with our proposed formulation under clean and contaminated distributions. Specifically, we investigate how our $L_{2,1}$-Flow Matching retains the performance of classical Flow Matching on clean data, whether it is more robust under contamination, and whether it depends on the proportion and type of contamination. The clean-target dataset we use is CIFAR-10 (Krizhevsky, 2009) and we consider several contamination mechanisms. We evaluate generation quality using the Fréchet Inception distance (FID) (Heusel et al., 2017).

Li & Dong (2026) perform similar studies on CIFAR-10 as we do and we adopt several contamination mechanisms used in their experiments. They use a different framework and report that they deliberately use a small model and limited training budget. We use a well-performing Flow Matching configuration instead and first verify that both regression objectives attain FID values comparable to the established result on clean CIFAR-10. This shows that the $L_{2,1}$ objective achieves generation quality comparable to standard Flow Matching on clean data and allows us to study its robustness in a competitive Flow Matching setting. Because the modeling frameworks, training budgets, and evaluation procedures differ, the absolute FID values across the two studies are not directly comparable. A limitation of our results is that the theoretical results concern optimal vector fields at the population level. They do not account for estimation from finite data, approximation by neural networks, or optimization error. The extent to which the theoretical robustness properties carry over to trained models therefore remains unresolved. To perform the experiments, we use the U-Net architecture from Tong et al. (2024a;b)[1]. The difference between our method and classic flow matching is the norm training objective: squared Euclidean vs. $L_{2,1}$. All remaining components of

[1] `https://github.com/atong01/conditional-flow-matching`

Table 1: Contamination mechanisms used in the experiments. Images represented in $[-1, 1]^{3\times32\times32}$.

Contamination	Description	Parameters
Random noise	Replace the entire image by independent uniform noise	$U[-1, 1]$
Gaussian noise	Add independent Gaussian noise and clip to $[-1, 1]$	$\sigma = 0.15$
Salt and pepper	Replace selected pixels by either -1 or 1	$\rho = 0.05$
Patch noise	Replace a uniformly located square patch by uniform noise	8×8 patch
Inversion	Invert all normalized pixel values, $T(x) = -x$	–
Gaussian blur	Convolve the image with an isotropic Gaussian kernel	$\sigma = 1.5$, 11×11
Motion blur	Convolve the image with a horizontal motion-blur kernel	kernel size 7

Table 2: FID on clean CIFAR-10 (lower is better). Entries report the mean $\pm$ standard deviation over three seeds. The lower mean is shown in bold.

Objective	FID
Squared	3.70 ± 0.08
Unsquared	$\mathbf{3.50} \pm 0.01$

training and sampling are held fixed across experiments. Tong et al. (2024a) report a FID of 3.659 for independent FM on CIFAR-10. We use a standard Gaussian source distribution, linear interpolation and independent couplings. We use the same interpolation, optimizer, batch size and training budget. Specifically, models are trained for $400\,000$ gradient steps with Adam and a learning rate of 2×10^{-4}, batch size of 128, 5000 warm-up steps, gradient clipping at 1.0, and an EMA decay of 0.9999. We report the mean and standard deviation over three seeds. We fix the seeds $1, 2, 3$. Both methods use the same initialization seeds, and contamination is applied to the same images. FID is evaluated against the clean CIFAR-10 training set using $50\,000$ samples generated from the exponential moving average (EMA) model using the adaptive Dormand-Prince (DOPRI5) (Dormand & Prince, 1980) solver with tolerance 10^{-5}.

5.1 Contamination mechanisms

As discussed in Section 4 we use the contamination model

$$P^{\text{cont}} = (1 - \epsilon)P + \epsilon Q$$

where P denotes the CIFAR-10 distribution. We consider contamination proportions $\epsilon \in \{0.05, 0.1, 0.15\}$ and one clean reference setting where $\epsilon = 0$. The contamination models we are testing are described in Table 1 and an example visualization is included in the supplementary material in Figure 1. Theorem 4.3 suggests that our Flow Matching method should be particularly advantageous when the contaminating distribution is well separated from the clean target. Motivated by this observation, we consider contamination mechanisms with different degrees of severity and structure. These range from replacement by noise through global and localized pixel perturbations, to blur transformations. The contaminating distribution Q may either be a completely independent replacement distribution or induced by applying corruption to clean images. Several of our contamination models are taken from Li & Dong (2026). Note that they combine the contamination models into one mixture contamination distribution, while we evaluate them separately.

5.2 Results

The results for the clean experiments are presented in Table 2. Both the standard and our method achieve similarly low FID values, while our method yields a lower mean FID. This is close to the 3.659 baseline reported by Tong et al. (2024a).

Table 3 reports the FID for the contaminated experiments with $\epsilon \in \{0.05, 0.1, 0.15\}$. Our method performs better, across all proportions of contamination, for Random noise, Gaussian noise, Salt and Pepper, Patch noise and Inversion. The standard method performs better, again across all proportions of contamination, for the Gaussian blur and Motion blur.

Table 3: FID on clean CIFAR-10 (lower is better) after training under contamination. Entries report the mean $\pm$ standard deviation over three seeds. The lower mean within each pair is bold.

Contamination	Objective	$\epsilon = 0.05$	$\epsilon = 0.10$	$\epsilon = 0.15$
Random noise	Squared	9.69 ± 0.20	19.16 ± 0.33	29.73 ± 0.61
	Unsquared	$\mathbf{6.82} \pm 0.26$	$\mathbf{12.21} \pm 0.28$	$\mathbf{19.40} \pm 0.82$
Gaussian noise	Squared	3.79 ± 0.06	4.92 ± 0.17	6.72 ± 0.20
	Unsquared	$\mathbf{3.51} \pm 0.07$	$\mathbf{4.17} \pm 0.06$	$\mathbf{5.31} \pm 0.05$
Salt and pepper	Squared	4.45 ± 0.04	6.49 ± 0.17	9.44 ± 0.53
	Unsquared	$\mathbf{3.92} \pm 0.10$	$\mathbf{5.00} \pm 0.09$	$\mathbf{6.65} \pm 0.32$
Patch noise	Squared	3.98 ± 0.05	4.61 ± 0.07	5.49 ± 0.08
	Unsquared	$\mathbf{3.68} \pm 0.07$	$\mathbf{4.10} \pm 0.08$	$\mathbf{4.70} \pm 0.11$
Inversion	Squared	3.96 ± 0.10	4.84 ± 0.18	6.06 ± 0.15
	Unsquared	$\mathbf{3.66} \pm 0.07$	$\mathbf{4.37} \pm 0.06$	$\mathbf{5.52} \pm 0.13$
Gaussian blur	Squared	$\mathbf{4.58} \pm 0.07$	$\mathbf{6.18} \pm 0.10$	$\mathbf{8.84} \pm 0.25$
	Unsquared	5.09 ± 0.11	7.98 ± 0.09	11.85 ± 0.07
Motion blur	Squared	$\mathbf{4.45} \pm 0.06$	$\mathbf{6.13} \pm 0.11$	$\mathbf{8.48} \pm 0.13$
	Unsquared	4.71 ± 0.09	7.18 ± 0.22	10.40 ± 0.21

5.3 Discussion and Limitations

The results on clean data suggest that in this particular model, the modified Flow Matching objective can preserve generation quality. Lemma 3.2 gives a sufficient condition for this to hold. However, we cannot verify its assumptions for the CIFAR-10 distribution, although the empirical observations are consistent with it.

The pattern for the contaminated models suggests that the $L_{2,1}$-Flow Matching method is effective for certain kinds of contamination. In particular, in the models under which the $L_{2,1}$ method performs better, contamination consists of strong perturbations of pixels, individually or across the whole image. We suspect that these contamination models result in substantial deviations from the clean image distribution, under which the $L_{2,1}$ objective is beneficial. The standard method performs better for both Gaussian blur and Motion blur. A possible explanation is that the blurring we applied produces a contamination that is less separated from the original distribution. This hypothesis is partially supported by Theorem 4.3. We showed that under the stated assumptions, for bounded clean data with increasingly distant Dirac contamination, eventually the $L_{2,1}$ method will yield a better result, in the Wasserstein metric. Moreover, this resembles the intuition that the geometric median is generally more useful for contamination that is strongly separated from the original data. Nevertheless, we should be careful with this intuition. The intuition regarding the geometric median applies only partially, because the velocity distributions $\mathrm{Law}(X_1 - X_0 | X_t = x)$, which are the basis of the geometric median, have different mixture weights than the contaminated target distributions.

6 Conclusion

In this work we proposed an adaptation of the Flow Matching method, obtained by replacing the regression objective with an $L_{2,1}$ objective. We gave a characterization of the solution as a conditional geometric median and established conditions under which this changed objective yields the same result as the standard method. We then investigated the effect that contamination has on the vector fields and provided bounds on the influence of contamination. We have established improvement in the Wasserstein distance for our method in specific contamination settings under stated assumptions. Empirically, the method achieves comparable performance on clean CIFAR-10 and improves generation quality under several, though not all, contamination mechanisms. We do not establish general regularity or uniqueness results for trajectories induced by the $L_{2,1}$ formulation. Moreover, pointwise robustness of the velocity field does not by itself guarantee that the generated distribution is closer to the clean target. Extending our distributional guarantees beyond the specific contamination settings that we analyzed therefore remains an important direction for future research.

AI USE STATEMENT

In this work, we used generative AI tools to assist with formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in writing of proofs and implementing methods and proposing or refining hypotheses. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, designing or providing feedback on research methodology or experiments, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or to help interpret results. Assisting with translation is not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, creating or editing software code, summarizing or analyzing existing literature, sourcing/searching for information, editing a research paper to improve readability, identifying relevant literature, suggesting a structure for a research paper.

We have reviewed all AI-assisted work. In particular, we independently checked all AI-assisted mathematical claims and proof ideas. Moreover, we inspected the reviewed AI-assisted code and verified literature-related suggestions against their original sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We do not identify ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we build on an established experimental framework cited in the paper and use fixed random seeds in our experiments. Experimental setups are described in Section 5 and in the supplementary material in Section B.

REFERENCES

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

J.R. Dormand and P.J. Prince. A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980. ISSN 0377-0427. doi: 10.1016/0771-050X(80)90013-3. URL https://www.sciencedirect.com/science/article/pii/0771050X80900133.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJxgknCcK7.

J. B. S. Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417, 1948. ISSN 0006-3444. doi: 10.1093/biomet/35.3-4.414. URL https://doi.org/10.1093/biomet/35.3-4.414.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Artem Khrapov, Vadim Popov, Tasnima Sadekova, Assel Yermekova, and Mikhail Kudinov. Improving diffusion models's data-corruption resistance using scheduled Pseudo-Huber loss, 2024. URL https://arxiv.org/abs/2403.16728.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Lei Li and Yuexiao Dong. Robust diffusion models via divergence-induced weighted denoising, 2026. URL https://arxiv.org/abs/2606.22521.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. URL https://arxiv.org/abs/2412.06264.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Hendrik Lopuhaä and Peter Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19:229–248, 1991. doi: 10.1214/aos/1176347978.

P. Milasevic and G. R. Ducharme. Uniqueness of the Spatial Median. *The Annals of Statistics*, 15 (3):1332–1333, 1987. doi: 10.1214/aos/1176350511. URL https://doi.org/10.1214/aos/1176350511.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4): 2308–2335, 2015. ISSN 1350-7265. doi: 10.3150/14-bej645. URL http://dx.doi.org/10.3150/14-BEJ645.

Stanislav Minsker and Nate Strawn. The geometric median and applications to robust mean estimation. *SIAM Journal on Mathematics of Data Science*, 6(2):504–533, 2024. doi: 10.1137/23M1592420. URL https://doi.org/10.1137/23M1592420.

Gabriel Peyré. Optimal and diffusion transports in machine learning. In *Proceedings of the International Congress of Mathematicians 2026 - Volume 7: Invited Lectures (Sections 15–20)*, pp. 110–128. Society for Industrial and Applied Mathematics, 2026. doi: 10.1137/25M1803966. URL https://epubs.siam.org/doi/abs/10.1137/25M1803966.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Peter Rousseeuw and Annick Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 1987. doi: 10.1002/0471725382.

Richard Schwank, Andrew McCormack, and Mathias Drton. Robust score matching. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1234–1242. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/schwank25a.html.

Christopher G. Small. A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277, 1990. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403809.

Weixi Song, Zhetao Chen, Tao Xu, Xianchao Zeng, Xinyu Zhou, Lixin Yang, Donglin Wang, Cewu Lu, and Yong-Lu Li. L1 sample flow for efficient visuomotor learning, 2025. URL https://arxiv.org/abs/2511.17898.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW.

Alexander Y. Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free Schrödinger bridges via score and flow matching. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1279–1287. PMLR, 02–04 May 2024b. URL https://proceedings.mlr.press/v238/tong24a.html.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

SUPPLEMENTARY MATERIAL

A PROOFS

Here we prove the statements from the main text. Throughout this appendix, we retain the notation and assumptions introduced in the main text. For convenience, we reintroduce the statements we prove.

Lemma (3.1)**.** *Let $v_t^1(x)$ be a measurable minimizer of Equation 3. Then it holds that*

$$v_t^1(x) \in \text{Gmed}(X_1 - X_0 | X_t = x),$$

for almost every t and P_t-almost every x. Moreover, if this conditional distribution is not concentrated on an affine line for almost every t and P_t-almost every x, then the minimizer is unique up to almost everywhere equality.

Proof. Let $K_t(x, \cdot)$ denote a jointly measurable probability kernel representing the conditional law of $X_1 - X_0$ given $X_t = x$. Then

$$\begin{aligned}\mathcal{L}(v) &= \int_0^1 \mathbb{E}[\|v_t(X_t) - (X_1 - X_0)\|]\mathrm{d}t \\ &= \int_0^1 \int_{\mathbb{R}^n \times \mathbb{R}^n} \|v_t(x) - z\| K_t(x, \mathrm{d}z) P_t(\mathrm{d}x)\mathrm{d}t.\end{aligned}$$

For fixed t, x the inner integral

$$r(t, x, a) \coloneqq \int_{\mathbb{R}^n} \|a - z\| K_t(x, \mathrm{d}z)$$

is minimized exactly when a is a geometric median of $K_t(x, \cdot)$. This holds for almost every t, x. Assume that $v_t^1(x)$ would not minimize $r(t, x, \cdot)$ on a set that has positive measure that we denote by E. By continuity in the last argument and since $\mathbb{Q}^n$ is dense there has to exist a rational vector $q^* \in \mathbb{Q}^n$ such that

$$r(t, x, q^*) < r(t, x, v_t^1(x))$$

for (t, x) in a set of positive measure. To see that, let

$$A_q \coloneqq \{(t, x) : r(t, x, q) < r(t, x, v_t^1(x))\}.$$

Since r is continuous in its last argument and $\mathbb{Q}^n$ is dense we can write

$$E = \bigcup_{q \in \mathbb{Q}^n} A_q.$$

Since E has positive measure, there has to exist a q^* such that A_{q^*} has positive measure. Consequently, replacing $v_t^1(x)$ with q^* on A_{q^*} would decrease $\mathcal{L}$, contradicting minimality. The uniqueness claim follows immediately from Milasevic & Ducharme (1987). □

Lemma (3.2)**.** *Assume that, for almost every t and P_t-almost every x, the conditional distribution of $X_1 - X_0$ given $X_t = x$ is not concentrated on an affine line and is centrally symmetric around its mean, i.e., for any random variable having this distribution, say Y, we have that $Y - \mathbb{E}[Y]$ matches, in distribution, $-(Y - \mathbb{E}[Y])$. Then we have that*

$$v_t^1(x) = v_t^2(x)$$

for almost every t and P_t-almost every x.

Proof. This holds for general distributions. Let μ be a distribution satisfying the assumptions above. Let

$$m = \int x \mu(\mathrm{d}x)$$

be the expectation of μ. Consider

$$F(y) = \int \|y - x\| \mu(\mathrm{d}x).$$

Since μ has finite expectation, the minimizer of F is the geometric median, which is unique by assumption. Since μ is assumed to be centrally symmetric around its mean, it holds that

$$F(y) = F(2m - y).$$

Since F is convex

$$F(m) = F\left(\frac{y + (2m - y)}{2}\right) \leq \frac{F(y) + F(2m - y)}{2} = F(y),$$

implying that m is the unique geometric median. Applying this result to the conditional distribution of $X_1 - X_0$ given $X_t = x$ and using the characterizations of $v_t^1(x)$ and $v_t^2(x)$ proves the claim. □

Corollary (3.3)**.** *If the source distribution is a standard Gaussian distribution $X_0 \sim \mathcal{N}(0, \mathrm{I})$ and the target distribution is nondegenerate Gaussian $X_1 \sim \mathcal{N}(\mu, \Sigma)$ where X_0 and X_1 are independent, then*

$$v_t^1(x) = v_t^2(x)$$

for almost every t and P_t-almost every x.

Proof. Since X_0 and X_1 are independent Gaussian random variables it follows that $X_t = (1 - t)X_0 + tX_1$ is Gaussian. As both $X_1 - X_0$ and X_t are linear combinations of X_0 and X_1, the pair $(X_1 - X_0, X_t)$ is jointly Gaussian. Thus $X_1 - X_0$ given $X_t = x$ is nondegenerate Gaussian. The statement then follows since Gaussian distributions have a unique geometric median equal to their mean. □

Theorem (3.4)**.** *Let P_t be the law of X_t. Under the standing assumptions, we have that:*

$$\int_0^1 \int_{\mathbb{R}^n} \|v_t^1(x) - v_t^2(x)\|^2 P_t(\mathrm{d}x)\mathrm{d}t \leq \int_0^1 \mathbb{E}\|(X_1 - X_0) - v_t^2(X_t)\|^2 \mathrm{d}t$$

where the right-hand side is the optimal squared FM regression loss.

Proof. The following inequalities hold for almost every t and P_t-almost every x. Let

$$v_t^2(x) = \mathbb{E}[X_1 - X_0 | X_t = x]$$

and

$$v_t^1(x) \in \mathrm{Gmed}[X_1 - X_0 | X_t = x].$$

Then

$$\|v_t^2(x) - v_t^1(x)\| = \|\mathbb{E}[(X_1 - X_0) - v_t^1(x)|X_t = x]\| \leq \mathbb{E}[\|(X_1 - X_0) - v_t^1(x)\||X_t = x].$$

Since $v_t^1(x)$ is a geometric median of the conditional distribution, it minimizes the conditional expected Euclidean distance. Consequently

$$\|v_t^2(x) - v_t^1(x)\| \leq \mathbb{E}[\|(X_1 - X_0) - v_t^1(x)\||X_t = x] \leq \mathbb{E}[\|(X_1 - X_0) - v_t^2(x)\||X_t = x].$$

Moreover

$$\mathbb{E}[\|(X_1 - X_0) - v_t^2(x)\||X_t = x] \leq \sqrt{\mathbb{E}[\|(X_1 - X_0) - v_t^2(x)\|^2|X_t = x]}.$$

Consequently

$$\|v_t^2(x) - v_t^1(x)\|^2 \leq \mathbb{E}[\|(X_1 - X_0) - v_t^2(x)\|^2|X_t = x]$$

and integrating with respect to P_t yields

$$\int_0^1 \int_{\mathbb{R}^n} \|v_t^1(x) - v_t^2(x)\|^2 P_t(\mathrm{d}x)\mathrm{d}t \leq \int_0^1 \mathbb{E}\|(X_1 - X_0) - v_t^2(X_t)\|^2 \mathrm{d}t$$

□

Lemma (4.1)**.** *For almost every t and P_t^{cont}-almost every x satisfying $\gamma_t(x) < 1/2$, there exists $C_t(x) \geq 0$, depending only on the clean model and not on the contamination, such that*

$$\|v_t^{1,\epsilon}(x) - v_t^1(x)\| \leq \frac{2(1 - \gamma_t(x))}{1 - 2\gamma_t(x)} C_t(x).$$

Proof. Fix t and x. We have shown in Equation 4 that we can decompose the velocity distribution as

$$P_{t,x}^{\text{cont}} = (1 - \gamma_t(x))P_{t,x} + \gamma_t(x)Q_{t,x}.$$

Let $F(z) = \int \|y - z\| P_{t,x}(\mathrm{d}y)$ and $G(z) = \int \|y - z\| Q_{t,x}(\mathrm{d}y)$. By definition $v_t^1(x)$ minimizes F, because $P_{t,x}$ is the clean conditional velocity distribution. On the other hand, by the decomposition of $P_{t,x}^{\text{cont}}$, it holds that $v_t^{1,\epsilon}(x)$ minimizes

$$F_{\text{cont}}(z) = (1 - \gamma_t(x))F(z) + \gamma_t(x)G(z).$$

Denote $v_t^1(x) = m$ and $v_t^{1,\epsilon}(x) = m^*$. Due to the optimality mentioned above

$$(1 - \gamma_t(x))F(m^*) + \gamma_t(x)G(m^*) \leq (1 - \gamma_t(x))F(m) + \gamma_t(x)G(m). \tag{6}$$

Let $d = \|m^* - m\|$.

Using the reverse triangle inequality we have

$$\|y - m^*\| \geq \|m^* - m\| - \|y - m\|$$

and

$$\|y - m^*\| \geq \|y - m\| - \|m^* - m\|.$$

By taking the appropriate expectations this implies

$$F(m^*) \geq d - F(m)$$

and

$$G(m^*) \geq G(m) - d.$$

Plugging those into Equation 6:

$$(1 - \gamma_t(x))(d - F(m)) + \gamma_t(x)(G(m) - d) \leq (1 - \gamma_t(x))F(m) + \gamma_t(x)G(m)$$

after reformulating

$$(1 - 2\gamma_t(x))d - (1 - \gamma_t(x))F(m) + \gamma_t(x)G(m) \leq (1 - \gamma_t(x))F(m) + \gamma_t(x)G(m)$$

equivalent to

$$(1 - 2\gamma_t(x))d \leq 2(1 - \gamma_t(x))F(m).$$

Since by assumption $\gamma_t(x) < 1/2$ it holds that $1 - 2\gamma_t(x) > 0$ and we conclude

$$d \leq \frac{2(1 - \gamma_t(x))}{1 - 2\gamma_t(x)} F(m).$$

With

$$C_t(x) = \int \|y - v_t^1(x)\| P_{t,x}(\mathrm{d}y)$$

the claim follows. □

Theorem (4.2). *Let δ_0 be a Dirac distribution in $\mathbb{R}^n$ centered at 0. Let $0 < \epsilon < 1/2$. For $b \in \mathbb{R}^n \setminus \{0\}$, consider the contaminated distribution*

$$P^{\text{cont}} = (1-\epsilon)\delta_0 + \epsilon\delta_b.$$

For every pointwise $L_{2,1}$-velocity field $v^{1,P^{\text{cont}}}$ coming from a Borel geometric-median selection as defined in Equation 5, the associated ODE admits a solution process $(Z_t)_{t\in[0,1)}$ with $Z_0 \sim \mathcal{N}(0, \mathrm{I}_n)$. Every such solution converges almost surely as $t \nearrow 1$ and its terminal law $\rho_1 = \mathrm{Law}(\lim_{t\nearrow 1} Z_t)$ satisfies

$$W_1(\rho_1, \delta_0) < W_1(P^{\text{cont}}, \delta_0).$$

Proof. As described in the main text, we will condition on the endpoint being either the clean or the contaminated target. We consider $0 \le t < 1$. Conditioned on either the clean or the contaminated endpoint, the distribution of the interpolation is Gaussian

$$X_t|(X_1 = \xi) \sim \mathcal{N}(t\xi, (1-t)^2\,\mathrm{I}) \tag{7}$$

where $\xi \in \{0, b\}$. Define the posterior contamination probability $\gamma_t(x) = \mathbb{P}[X_1 = b|X_t = x]$ and

$$\Gamma_t(x) = \frac{\gamma_t(x)}{(1-\gamma_t(x))} = \frac{\mathbb{P}[X_1 = b|X_t = x]}{\mathbb{P}[X_1 = 0|X_t = x]}.$$

We want to analyze this quantity, since it explains the behavior of the $L_{2,1}$ field. From Equation 7 we obtain that $X_t|(X_1 = \xi)$ admits a density. Using $\mathbb{P}[X_1 = b] = \epsilon$ and the law of total probability we obtain that for the density $p_t(x)$ of X_t the following holds:

$$p_t(x) = (1-\epsilon)p_t(x|X_1 = 0) + \epsilon p_t(x|X_1 = b).$$

Consequently using Bayes' theorem

$$\gamma_t(x) = \frac{\epsilon p_t(x|X_1 = b)}{(1-\epsilon)p_t(x|X_1 = 0) + \epsilon p_t(x|X_1 = b)}.$$

Similarly,

$$1 - \gamma_t(x) = \frac{(1-\epsilon)p_t(x|X_1 = 0)}{(1-\epsilon)p_t(x|X_1 = 0) + \epsilon p_t(x|X_1 = b)}.$$

For $\Gamma_t(x)$ the common denominator cancels and it follows that

$$\Gamma_t(x) = \frac{\epsilon}{1-\epsilon}\frac{p_t(x|X_1 = b)}{p_t(x|X_1 = 0)}.$$

With the characterization of these densities in Equation 7 we obtain

$$\Gamma_t(x) = \frac{\epsilon}{1-\epsilon}\exp\left(-\frac{\|x - tb\|^2 - \|x\|^2}{2(1-t)^2}\right).$$

Taking the logarithm yields

$$\log\Gamma_t(x) = \log\frac{\epsilon}{1-\epsilon} + \frac{t\langle x, b\rangle}{(1-t)^2} - \frac{t^2\|b\|^2}{2(1-t)^2}. \tag{8}$$

Conditioned on an endpoint and the interpolation point, the source point is deterministic. Conditioning on $X_t = x$ and $X_1 = \xi$ gives

$$X_0 = \frac{x - t\xi}{1-t}, \quad X_1 - X_0 = \xi - X_0 = \frac{\xi - x}{1-t}.$$

For each $\xi \in \{0, b\}$ define the conditional field

$$u_t^\xi(x) = \frac{\xi - x}{1-t}.$$

In the sense of probability laws we have

$$\mathrm{Law}(X_1 - X_0|X_t = x) = (1-\gamma_t(x))\delta_{u_t^0(x)} + \gamma_t(x)\delta_{u_t^b(x)}.$$

For fixed $t < 1$ and x, this conditional law is a two-point mixture. If $\gamma_t(x) \neq 1/2$, its geometric median is the atom with higher assigned probability. Note that $\gamma_t(x) < 1/2 \iff \Gamma_t(x) < 1 \iff \log \Gamma_t(x) < 0$. Fix any Borel geometric-median selection $m^{P^{\text{cont}}}$ as defined in the main text and write

$$v_t^1(x) = v_t^{1,P^{\text{cont}}}(x) = \frac{m_t^{P^{\text{cont}}}(x) - x}{1-t}$$

Since the relevant conditional law is a two-point mixture as discussed, the field satisfies

$$v_t^1(x) = \begin{cases} u_t^0(x) & \text{if } \Gamma_t(x) < 1 \\ (1-\lambda_t(x))u_t^0(x) + \lambda_t(x)u_t^b(x) & \text{if } \Gamma_t(x) = 1 \\ u_t^b(x) & \text{if } \Gamma_t(x) > 1, \end{cases}$$

where $\lambda_t(x) \in [0,1]$ is determined by the fixed Borel selection, which we leave arbitrary. Note that this is relevant since for $\gamma_t(x) = 1/2$ both mixture weights are equal, thus the set of geometric medians is the line segment $[u_t^0(x), u_t^b(x)]$. For fixed source points, we can analyze the trajectories induced by the field. For a starting point x_0, the trajectory induced by the uncontaminated field is $T_t(x_0) = (1-t)x_0$. Let $c = -\log \frac{\epsilon}{1-\epsilon} > 0$ and $S(x_0) = \frac{\langle b, x_0 \rangle}{\|b\|}$. Using Equation 8:

$$\log \Gamma_t(T_t(x_0)) = -c + \frac{t}{1-t}\|b\|S(x_0) - \frac{1}{2}\frac{t^2}{(1-t)^2}\|b\|^2.$$

Let $z = \frac{t}{(1-t)}$. Then this is a concave quadratic in z:

$$\log \Gamma_t(T_t(x_0)) = -c + \|b\|S(x_0)z - \frac{1}{2}\|b\|^2 z^2 =: G_{x_0}(z).$$

Since the transformation $t \mapsto z$ is a continuous increasing function from $[0,1)$ onto $[0,\infty)$ it suffices to study the time-reparametrized quadratic G_{x_0}. If $S(x_0) > 0$ then its maximum over $z \geq 0$ is $-c + \frac{1}{2}S(x_0)^2$, while for $S(x_0) \leq 0$ its maximum is $-c < 0$. Consequently, if

$$S(x_0) < \sqrt{2c},$$

then

$$\log \Gamma_t(T_t(x_0)) < 0,$$

for every $t < 1$. We are interested in the sign of G_{x_0}, since

$$G_{x_0}(z) < 0 \iff \Gamma_{\frac{z}{1+z}}(T_{\frac{z}{1+z}}(x_0)) = \Gamma_t(T_t(x_0)) < 1.$$

Define $H_{\text{clean}} = \{x_0 \in \mathbb{R}^n : S(x_0) < \sqrt{2c}\}$ and let $x_0 \in H_{\text{clean}}$. By what we have shown above, $G_{x_0}(z) < 0$ for every $z \geq 0$. In particular this implies that the vector field selects the clean velocity along the candidate clean trajectory

$$v_t^1(T_t(x_0)) = u_t^0(T_t(x_0)) = \frac{-T_t(x_0)}{1-t}.$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}t}T_t(x_0) = -x_0 = \frac{-T_t(x_0)}{1-t}$$

the clean trajectory is indeed a trajectory of the $L_{2,1}$-field ODE on $[0,1)$. Note that Γ is continuous and the trajectory $T_t(x_0)$ is also continuous. Since $\Gamma_t(T_t(x_0)) < 1$, for every $t < 1$ there exists a neighborhood around $(t, T_t(x_0))$ such that $\Gamma < 1$. Consequently the geometric median field agrees with

$$u_s^0(y) = -\frac{y}{1-s}$$

for all s, y in that neighborhood. This field is locally Lipschitz in space and hence the clean trajectory is indeed the unique trajectory starting from x_0.

Hence, starting points $x_0 \in H_{\text{clean}}$ follow $T_t(x_0)$ and consequently get transported to the clean target since $T_t(x_0) = (1-t)x_0 \to 0$ for $t \nearrow 1$.

Now define $H_{\text{cont}} = \{x_0 \in \mathbb{R}^n : S(x_0) > \sqrt{2c}\}$ and let $x_0 \in H_{\text{cont}}$. Consider $G_{x_0}(z)$ as defined above. Since $S(x_0) > \sqrt{2c}$ it has a positive maximum and consequently has the two positive roots

$$z_\pm = \frac{S(x_0) \pm \sqrt{S(x_0)^2 - 2c}}{\|b\|}.$$

Consider $z_- = \frac{S(x_0) - \sqrt{S(x_0)^2 - 2c}}{\|b\|}$ and $t_- = \frac{z_-}{1+z_-}$. Consider any locally absolutely continuous trajectory with initial value x_0, solving the flow ODE for almost every t. Analogously to the previous argument, since $G_{x_0}(z) < 0$ for $0 \le z < z_-$ the trajectory initially follows the clean field for $0 \le t < t_-$. After a potential switch, the candidate trajectory is $T_t(x_0) = (1-t)Y(z)$ where $Y(z) = x_0 + (z - z_-)b$. Its posterior log-odds are

$$L_{x_0}(z) = -c + \langle Y(z), b\rangle z - \frac{1}{2}\|b\|^2 z^2$$

with $L_{x_0}(z_-) = 0$ and since $Y'(z) = b$ we have

$$L'_{x_0}(z) = \langle Y(z), b\rangle = \|b\| S(x_0) + (z - z_-)\|b\|^2 > 0 \quad \text{for every } z > z_-,$$

implying $L_{x_0}(z) > 0$ for every $z > z_-$. Note that $G'_{x_0}(z_-) > 0$ and $L'_{x_0}(z_-) > 0$. Since $dz/dt = (1-t)^{-2} > 0$, the corresponding time derivatives are also positive at the switching point. By a continuity argument analogous to the one above, the derivatives of the log-odds along both branch velocities remain bounded below by a common positive constant in a neighborhood of the switching point. On the tie surface, the velocity is a convex combination of $u_t^0(x)$ and $u_t^b(x)$. Write $l(t,x) = \log \Gamma_t(x)$. For the selected velocity, the chain rule gives

$$\partial_t l + \langle \nabla_x l, v_t^1\rangle = (1-\lambda)(\partial_t l + \langle \nabla_x l, u_t^0\ \rangle) + \lambda(\partial_t l + \langle \nabla_x l, u_t^b\ \rangle) \ge \eta > 0,$$

where $\lambda \in [0,1]$ is the coefficient of the convex combination and η is the common positive lower bound. All quantities are evaluated at (t,x). Hence, along any absolutely continuous solution through this point, the log-odds increase strictly for a sufficiently short time, so the solution immediately enters the contaminated region. Local uniqueness there identifies it with the candidate trajectory, which remains in that region since $L_{x_0}(z) > 0$ for every $z > z_-$.

Hence, the trajectory after z_- is $T_t(x_0) = (1-t)Y(z) = \frac{x_0 + (z - z_-)b}{1+z}$. As $t \nearrow 1$, $z \to \infty$ so the trajectory converges to b. This means that initial points $x_0 \in H_{\text{cont}}$ get transported to b.

Putting this together, the trajectories follow

$$T_t(x_0) = \begin{cases} (1-t)x_0 & \text{if } S(x_0) \le \sqrt{2c} \\ (1-t)x_0 + \max\{0, t - (1-t)z_-(x_0)\}b & \text{if } S(x_0) > \sqrt{2c}. \end{cases}$$

This map is Borel measurable and the paths are locally absolutely continuous. Let $X_0 \sim \mathcal{N}(0, \mathrm{I}_n)$ and define

$$Z_t \coloneqq T_t(X_0), \quad 0 \le t < 1.$$

This establishes the existence of a solution process. Our preceding calculations show that for every x_0 satisfying $S(x_0) \ne \sqrt{2c}$, $T(x_0)$ satisfies the selected ODE almost everywhere. Since $S(X_0) \sim \mathcal{N}(0,1)$ this condition holds almost surely. Moreover, the preceding uniqueness arguments show that every solution starting from x_0 with $S(x_0) \ne \sqrt{2c}$ coincides with $T_t(x_0)$. Consequently, every solution process with the stated Gaussian initial distribution agrees almost surely, for all $t < 1$, with $T_t(Z_0)$. The previously established trajectory limits therefore give almost-sure convergence and the same terminal law for every such process.

We want to analyze the induced terminal distribution. Since $X_0 \sim \mathcal{N}(0, \mathrm{I})$ it follows that

$$S(X_0) = \left\langle \frac{b}{\|b\|}, X_0 \right\rangle \sim \mathcal{N}(0,1)$$

since S is a normalized linear projection. Thus

$$\epsilon^* = \mathbb{P}[S(X_0) > \sqrt{2c}] = 1 - \Phi\left(\sqrt{2c}\right).$$

Consequently, the induced distribution is

$$\rho_1 = (1-\epsilon^*)\delta_0 + \epsilon^*\delta_b,$$

from which the Wasserstein statement follows immediately since for $a \in [0,1]$:

$$W_1((1-a)\delta_0 + a\delta_b, \delta_0) = a\|b\|.$$

Then

$$\epsilon^* = 1 - \Phi\left(\sqrt{2c}\right) < \frac{1}{2}e^{-c} = \frac{\epsilon}{2(1-\epsilon)} < \epsilon,$$

concluding the proof. □

Theorem (4.3)**.** *Let P be a distribution supported in the closed ball $\overline{B}_1(0)$. Let $0 < \epsilon < 1/2$, and fix a unit vector $u \in \mathbb{R}^n$. For $R > 1$, set $b_R = Ru$ and*

$$P_R^{\text{cont}} = (1-\epsilon)P + \epsilon\delta_{b_R}.$$

For each $R > 1$, let $v^{1,R}$ be any pointwise $L_{2,1}$-velocity field associated with P_R^{cont} as defined in Equation 5. Assume that its associated ODE admits a solution process with initial distribution $\mathcal{N}(0, I_n)$. Let $(Z_t^R)_{t\in[0,1)}$ be any such solution process, and write $\rho_t^R = \text{Law}(Z_t^R)$. Then, for all sufficiently large R,

$$\limsup_{t\nearrow 1} W_1(\rho_t^R, P) < W_1(P_R^{\text{cont}}, P).$$

Proof. For each $R > 1$, let m^R be the Borel geometric-median selection defining $v^{1,R}$, and let $(Z_t^R)_{t\in[0,1)}$ be any associated solution process as in the theorem. We suppress the dependence on R in m^R and Z_t^R. By the definition of a solution, almost surely,

$$Z_t = Z_0 + \int_0^t \frac{m_s(Z_s) - Z_s}{1-s}\,\mathrm{d}s, \qquad \text{for every } 0 \le t < 1. \tag{9}$$

The beginning of this proof is analogous to the proof of Theorem 4.2. First we condition on the endpoint. For $0 \le t < 1$ it holds that

$$X_t | (X_1 = \xi) \sim \mathcal{N}(t\xi, (1-t)^2\,\mathrm{I}), \tag{10}$$

where $\xi \in \overline{B}_1(0) \cup \{b_R\}$. Since those again admit densities we can write

$$l_\xi(t,x) \coloneqq \exp\left(\frac{-\|x - t\xi\|^2}{2(1-t)^2}\right).$$

The same Bayes computation as before yields

$$\Gamma_t(x) \coloneqq \frac{\gamma_t(x)}{(1-\gamma_t(x))} = \frac{\epsilon}{1-\epsilon}\frac{l_{b_R}(t,x)}{\int l_\xi(t,x)P(\mathrm{d}\xi)}. \tag{11}$$

The main difference is that in the denominator we have to integrate over the target distribution, since it is no longer a point mass distribution. The clean posterior is $\nu_{t,x}(\mathrm{d}y) = \frac{l_y(t,x)P(\mathrm{d}y)}{\int l_\xi(t,x)P(\mathrm{d}\xi)}$. For this distribution it holds that $\text{supp}\,\nu_{t,x} \subset \overline{B}_1(0)$. The conditional distribution can now be characterized as

$$\text{Law}(X_1 | X_t = x) = (1-\gamma_t(x))\nu_{t,x} + \gamma_t(x)\delta_{b_R}.$$

By the definition of the fixed selection m, we therefore have

$$m_t(x) \in \text{Gmed}\left((1-\gamma_t(x))\nu_{t,x} + \gamma_t(x)\delta_{b_R}\right)$$

for every $(t,x) \in [0,1) \times \mathbb{R}^n$.

In the earlier proof, we could give a precise formula for the geometric median, since the velocity distribution consisted of two atoms. Since we are working with a general bounded target distribution, we will instead provide a bound for the geometric median. Fix a threshold $\epsilon < \alpha < 1/2$ and introduce constants

$$C_0 = \frac{2(1-\alpha)}{1-2\alpha}, \quad L_0 = \log\left(\frac{\alpha}{1-\alpha}\frac{1-\epsilon}{\epsilon}\right) > 0.$$

Along the trajectory define

$$\tau = \inf\{t \in [0,1) : \gamma_t(Z_t) \geq \alpha\}$$

with the convention that the infimum of the empty set is 1. This is the first time that the posterior contamination probability along the trajectory reaches the threshold. Let

$$H_R = \{\tau < 1\}$$

and H_R^c its complement. On the event H_R^c we have that $\tau = 1$ and thus

$$\gamma_t(Z_t) < \alpha \qquad \text{for every } 0 \leq t < 1.$$

This allows us to bound the median by C_0 using Lemma A.1. Independently of H_R, since the clean conditional endpoint distribution is supported in $\overline{B}_1(0)$, the mixed conditional endpoint distribution is supported in $\overline{B}_1(0) \cup \{b_R\} \subset \overline{B}_R(0)$. Note that the geometric median remains inside $\overline{B}_R(0)$ because otherwise projecting it onto this ball would strictly decrease its distance to every support point. Thus

$$\|m_t(Z_t)\| \leq R.$$

In particular, this yields the following bound:

$$\|m_r(Z_r)\| \leq C_0 \mathbf{1}_{H_R^c} + R\mathbf{1}_{H_R}. \tag{12}$$

Now we want to use the bound on the median to provide a bound on the resulting distribution. Let $\rho_t = \text{Law}(Z_t)$. Take $Y \sim P$ independent of the generated process. Then this yields an admissible coupling between ρ_t and P. Since the Wasserstein distance is the infimum of the transport cost of all admissible couplings, we can use this to bound the Wasserstein distance. We have

$$W_1(\rho_t, P) \leq \mathbb{E}\|Z_t - Y\| \leq \mathbb{E}\|Z_t\| + 1 \tag{13}$$

since $Y \in \overline{B}_1(0)$ almost surely. In the proof of Lemma A.2 we have shown

$$Z_t = (1-t)Z_0 + (1-t)\int_0^t \frac{m_r(Z_r)}{(1-r)^2}\mathrm{d}r. \tag{14}$$

Consequently

$$\mathbb{E}\|Z_t\| \leq (1-t)\mathbb{E}\|Z_0\| + (1-t)\int_0^t \frac{\mathbb{E}\|m_r(Z_r)\|}{(1-r)^2}\mathrm{d}r. \tag{15}$$

Using Equation 12 we obtain

$$\mathbb{E}\|m_t(Z_t)\| \leq (1 - \mathbb{P}(H_R))C_0 + \mathbb{P}(H_R)R.$$

Plugging this into Equation 15 and explicitly computing the integral yields

$$W_1(\rho_t, P) \leq \mathbb{E}\|Z_t\| + 1 \leq (1-t)\mathbb{E}\|Z_0\| + 1 + t((1 - \mathbb{P}(H_R))C_0 + \mathbb{P}(H_R)R).$$

Taking now $t \nearrow 1$ we obtain

$$\limsup_{t \nearrow 1} W_1(\rho_t, P) \leq C_0 + 1 + \mathbb{P}(H_R)(R - C_0).$$

Dividing by R and also sending R to infinity yields

$$\begin{aligned}\limsup_{R\to\infty} \frac{1}{R} \limsup_{t\nearrow 1} W_1(\rho_t, P) &\leq \limsup_{R\to\infty} \mathbb{P}(H_R) \\ &\leq (1 - \Phi(\sqrt{2L_0}))\end{aligned}$$

where the last inequality is by Lemma A.2. Since this inequality holds for all $\alpha \in (\epsilon, 1/2)$ and the right hand side is decreasing in α we can take the infimum over α and obtain

$$\limsup_{R\to\infty} \frac{1}{R} \limsup_{t\nearrow 1} W_1(\rho_t, P) \leq \lim_{\alpha \nearrow 1/2} (1 - \Phi(\sqrt{2L_0})) = 1 - \Phi\left(\sqrt{2\log\frac{1-\epsilon}{\epsilon}}\right).$$

Note that this is the same constant as in the proof of the Dirac contamination theorem. Moreover

$$\epsilon(R-1) \leq W_1(P_R^{\text{cont}}, P) \leq \epsilon(R+1),$$

implying

$$\frac{W_1(P_R^{\text{cont}}, P)}{R} \to \epsilon.$$

Since we have shown that

$$1 - \Phi\left(\sqrt{2\log\frac{1-\epsilon}{\epsilon}}\right) < \epsilon$$

the statement follows. □

Lemma A.1. *Consider the setting of Theorem 4.3. Let $0 \leq \gamma < 1/2$, let P be any distribution supported in $\overline{B}_1(0)$ and let m be any geometric median of*

$$(1-\gamma)P + \gamma\delta_b$$

for $b \in \mathbb{R}^n$. Then

$$\|m\| \leq \frac{2(1-\gamma)}{1-2\gamma}.$$

Proof. Note that this is a similar setup to Lemma 4.1 and a related result is (Minsker, 2015, Lemma 2.1). Let $F(z) = (1-\gamma)\int\|z-y\|P(\mathrm{d}y) + \gamma\|z-b\|$. Then m is its minimizer. Consequently $F(m) \leq F(0)$. Thus

$$F(m) - F(0) = (1-\gamma)\int(\|m-y\| - \|y\|)P(\mathrm{d}y) + \gamma(\|m-b\| - \|b\|)$$

implying that the right-hand side is non-positive. Consequently

$$\begin{aligned} 0 &\geq (1-\gamma)\int(\|m-y\| - \|y\|)P(\mathrm{d}y) + \gamma(\|m-b\| - \|b\|) \\ &\geq (1-\gamma)(\|m\| - 2) - \gamma\|m\| \\ &= (1-2\gamma)\|m\| - 2(1-\gamma). \end{aligned}$$

□

Lemma A.2. *In the setting of Theorem 4.3 and with the notation introduced in its proof, for every $\alpha \in (\epsilon, 1/2)$,*

$$\limsup_{R\to\infty} \mathbb{P}(H_R) \leq (1 - \Phi(\sqrt{2L_0}))$$

Proof. The function $C(\gamma) = 2(1-\gamma)/(1-2\gamma)$ is increasing on $[0, 1/2)$. For $t < \tau$, we have $\gamma_t(Z_t) < \alpha$, so Lemma A.1 yields

$$\|m_t(Z_t)\| \leq C(\gamma_t(Z_t)) \leq C(\alpha) = C_0.$$

By the assumptions of Theorem 4.3, the sample paths of Z are locally absolutely continuous and satisfy the integral equation almost surely. Therefore, for almost every $t \in [0, 1)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{Z_t}{1-t}\right) = \frac{m_t(Z_t)}{(1-t)^2}.$$

Integrating this from 0 to t yields

$$Z_t = (1-t)Z_0 + (1-t)\int_0^t \frac{m_r(Z_r)}{(1-r)^2}\mathrm{d}r.$$

For $0 < t < 1$ with $t \leq \tau$ define

$$c_t = \frac{1-t}{t}\int_0^t \frac{m_r(Z_r)}{(1-r)^2}\mathrm{d}r.$$

This yields

$$Z_t = (1-t)Z_0 + tc_t, \tag{16}$$

and using $\|m_t(Z_t)\| \le C_0$ we obtain that $\|c_t\| \le C_0$, which is useful because it is independent of R.

Now suppose H_R occurs.

The posterior formula is continuous as well as Z_t. Consequently, since $\gamma_0(Z_0) = \epsilon < \alpha$ we have $\gamma_\tau(Z_\tau) = \alpha$. Consider now $t = \tau\,, x = Z_\tau$. From the explicit posterior odds formula Equation 11 this yields

$$\exp(L_0) = \frac{l_{b_R}(t,x)}{\int l_\xi(t,x)P(\mathrm{d}\xi)} \le \sup_{\|y\|\le 1} \frac{l_{b_R}(t,x)}{l_y(t,x)},$$

where

$$L_0 = \log\left(\frac{\alpha(1-\epsilon)}{(1-\alpha)\epsilon}\right).$$

The inequality holds, since P is supported on the unit ball. Denote

$$s = 1 - t, z = t/s, S(x_0) = \langle x_0, u\rangle,$$

and take $y \in \overline{B}_1(0)$. From Equation 16 we obtain

$$x = sZ_0 + tc_t.$$

Denoting $d_y = b_R - y$ we obtain

$$\log \frac{l_{b_R}(t,x)}{l_y(t,x)} = z\langle Z_0, d_y\rangle + z^2\langle c_t - y, d_y\rangle - \frac{z^2}{2}\|d_y\|^2.$$

Using $\|y\| \le 1, \|c_t\| \le C_0$ and $b_R = Ru$ yields:

$$\langle Z_0, d_y\rangle \le RS(Z_0) + \|Z_0\|,\ \langle c_t - y, d_y\rangle \le M_0(R+1),\ \|d_y\| \ge R - 1.$$

where $M_0 = C_0 + 1$. Consequently, for all $\|y\| \le 1$ we have

$$\log \frac{l_{b_R}(t,x)}{l_y(t,x)} \le z\big(RS(Z_0) + \|Z_0\|\big) - \frac{z^2}{2}A_R.$$

where $A_R = (R-1)^2 - 2M_0(R+1)$. Now for all sufficiently large R it holds that $A_R > 0$. Combining the posterior-odds identity with the preceding uniform estimate yields

$$L_0 \le zB_R - \frac{z^2}{2}A_R,$$

where $B_R = RS(Z_0) + \|Z_0\|$. Since $L_0 > 0$ and for R large enough, it holds that $A_R > 0$, we have $B_R > 0$. Consequently

$$L_0 \le \max_{z\ge 0}\left(zB_R - \frac{A_R z^2}{2}\right) = \frac{B_R^2}{2A_R}.$$

This is equivalent to

$$S(Z_0) + \frac{\|Z_0\|}{R} \ge \frac{\sqrt{2L_0A_R}}{R}, \tag{17}$$

and is now independent of the hitting time τ. Let $G \sim \mathcal{N}(0, \mathrm{I}_n)$ be a Gaussian random variable independent of the other variables. Let

$$Y = \langle G, u\rangle \sim \mathcal{N}(0,1).$$

Since G has the same distribution as Z_0 and using Equation 17, the following holds:

$$\mathbb{P}(H_R) \le \mathbb{P}\left(Y + \frac{\|G\|}{R} \ge \frac{\sqrt{2L_0A_R}}{R}\right). \tag{18}$$

Note that

$$Y + \frac{\|G\|}{R} \to Y, \quad \frac{\sqrt{2L_0A_R}}{R} \to \sqrt{2L_0},$$

almost surely and

$$\mathbb{P}(Y = \sqrt{2L_0}) = 0.$$

Using the bounded convergence theorem on the indicator variables of the events yields

$$\lim_{R\to\infty} \mathbb{P}\left(Y + \frac{\|G\|}{R} \geq \frac{\sqrt{2L_0 A_R}}{R}\right) = 1 - \Phi(\sqrt{2L_0}).$$

Using that for sufficiently large R, Equation 18 holds we consequently have

$$\limsup_{R\to\infty} \mathbb{P}(H_R) \leq 1 - \Phi(\sqrt{2L_0}).$$

□

B Implementation Details

Our implementation is based on the CIFAR-10 code released in the TorchCFM repository accompanying Tong et al. (2024a;b). We use their CIFAR-10 I-CFM configuration and retain its U-Net architecture, data preprocessing, and principal training configuration. We modify the regression objective in the $L_{2,1}$ experiments and introduce the contamination mechanisms described in Table 1. CIFAR-10 images are converted to tensors, and normalized channelwise with mean and standard deviation 0.5, yielding images in $[-1, 1]^{3\times32\times32}$. For a batch of residuals $r_i = v_t(x_{t,i}) - u_{t,i}$, with $D = 3 \cdot 32 \cdot 32$, the squared loss is implemented as

$$\frac{1}{BD}\sum_{i=1}^{B}\|r_i\|_2^2,$$

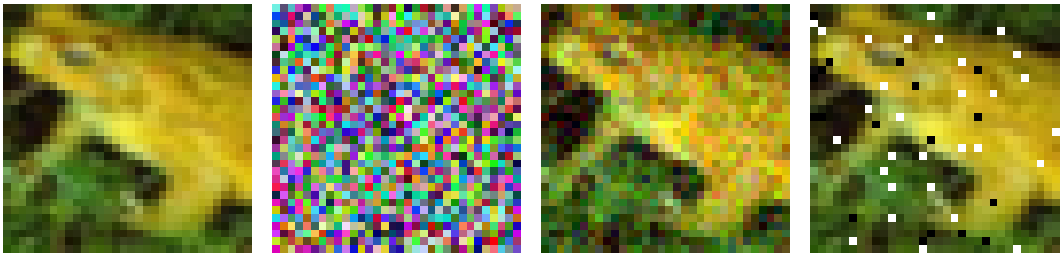
while the $L_{2,1}$ loss is implemented as

$$\frac{1}{B\sqrt{D}}\sum_{i=1}^{B}\|r_i\|_2.$$

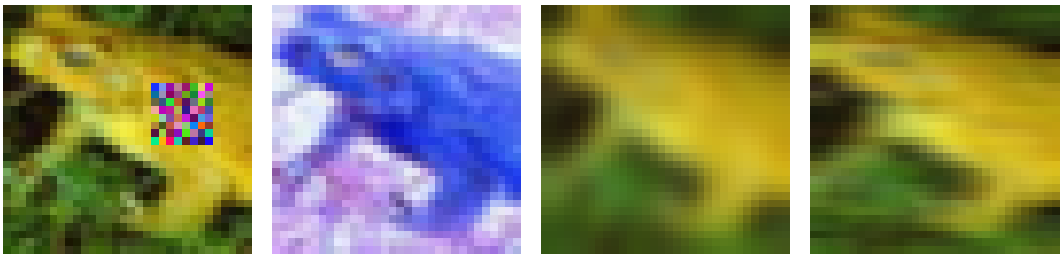
No smoothing is applied to the Euclidean norm. Since the dimension-dependent normalization factors are constant they do not change the population minimizers, but they can affect gradient clipping and are therefore reported as used in training. For each of the contamination experiments, the contamination is generated online: every target image occurring in a minibatch is independently corrupted with probability ϵ. Whether an image gets corrupted and associated noise are generated using a separate seeded random-number generator, so that matching randomization is used across objectives. We use a U-Net with 128 base channels, channel multipliers $[1, 2, 2, 2]$, attention at resolution 16, dropout 0.1, four attention heads and 64 channels per head. Models are trained for 400,000 optimizer updates using Adam with learning rate $2 \cdot 10^{-4}$, batch size 128, 5,000 linear warm-up steps, global gradient-norm clipping at 1.0, and EMA decay 0.9999.

For FID evaluation, we use the CIFAR-10 evaluation script provided in the TorchCFM repository accompanying Tong et al. (2024a;b), with only the modifications required to load and evaluate our trained models. We otherwise retain the original evaluation procedure and use the same pipeline for both the squared and $L_{2,1}$ objectives.

C EXAMPLE IMAGES

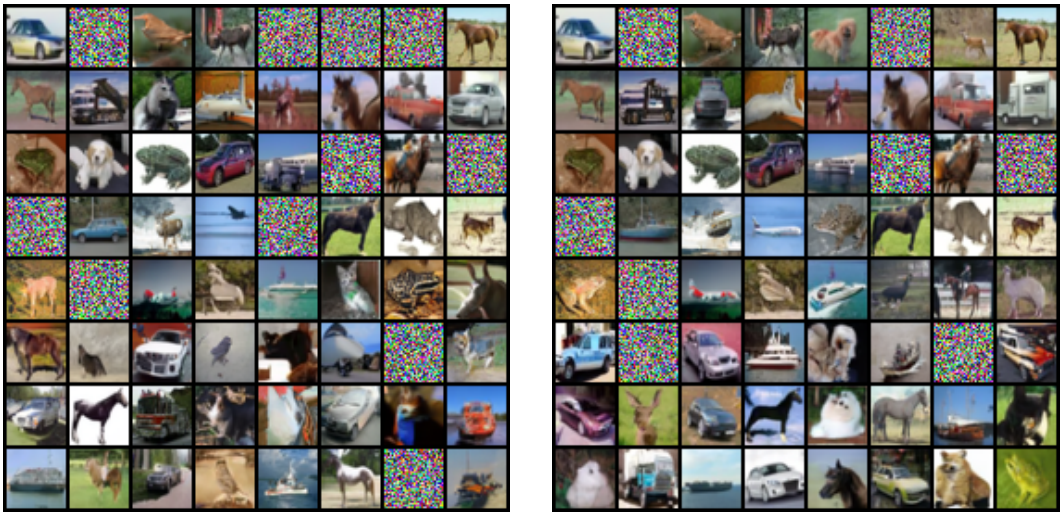


(a) Clean reference (b) Random noise (c) Gaussian noise (d) Salt-and-Pepper



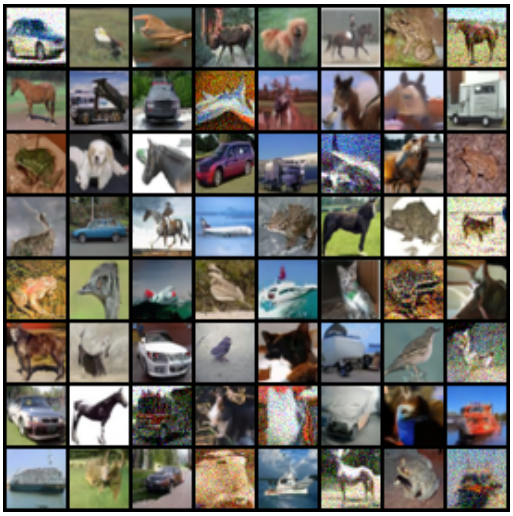
(e) Patch noise (f) Inversion (g) Gaussian blur (h) Motion blur

Figure 1: Representative examples of the seven contamination mechanisms considered in our experiments. Except for random-noise replacement, all examples are obtained from the same clean CIFAR-10 image using the parameters reported in Table 1.

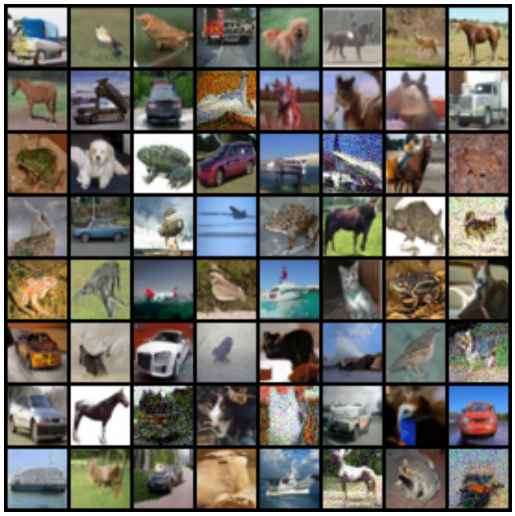


(a) Squared FM (b) $L_{2,1}$ FM

Figure 2: Generated samples for Random noise with contamination level $\epsilon = 0.15$. The same source samples are used for squared and $L_{2,1}$ Flow Matching.

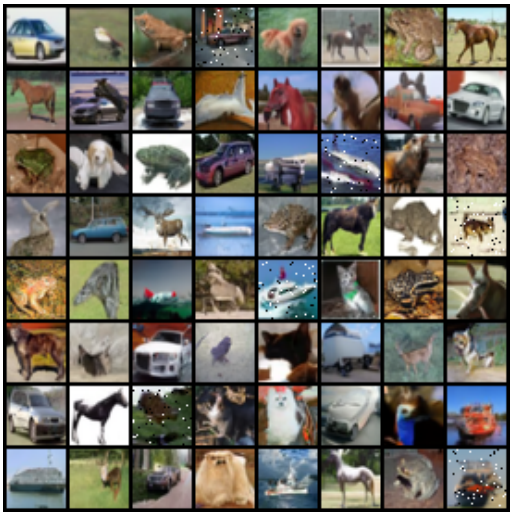


(a) Squared FM

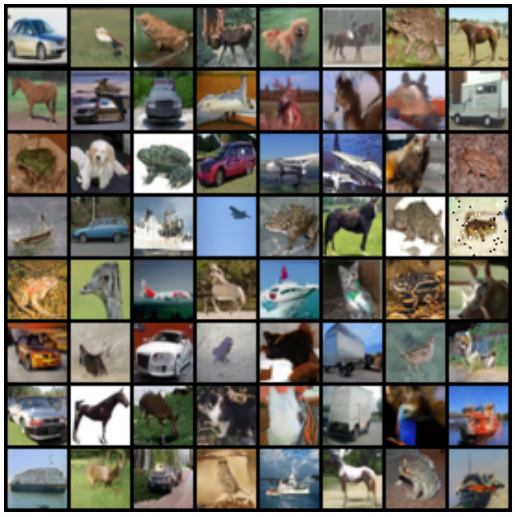


(b) $L_{2,1}$ FM

Figure 3: Generated samples for Gaussian noise with contamination level $\epsilon = 0.15$. The same source samples are used for squared and $L_{2,1}$ Flow Matching.

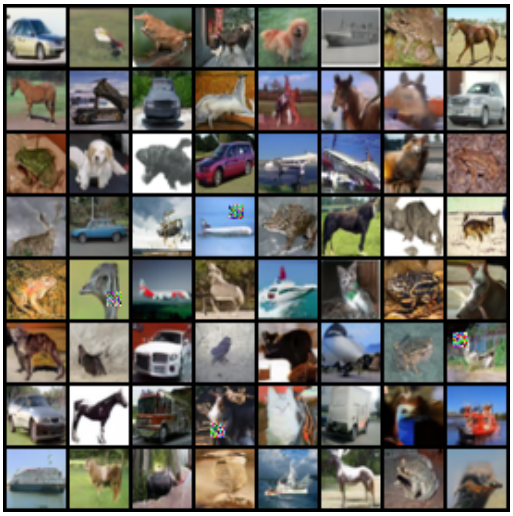


(a) Squared FM

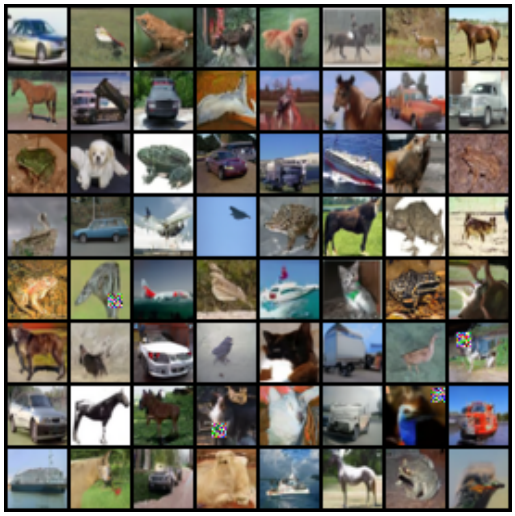


(b) $L_{2,1}$ FM

Figure 4: Generated samples for Salt-and-Pepper noise with contamination level $\epsilon = 0.15$. The same source samples are used for squared and $L_{2,1}$ Flow Matching.

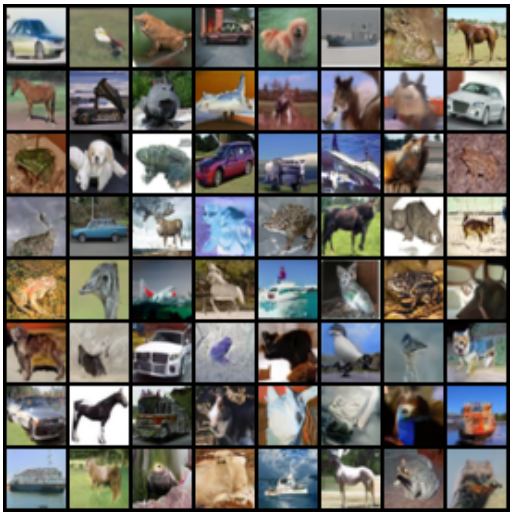


(a) Squared FM

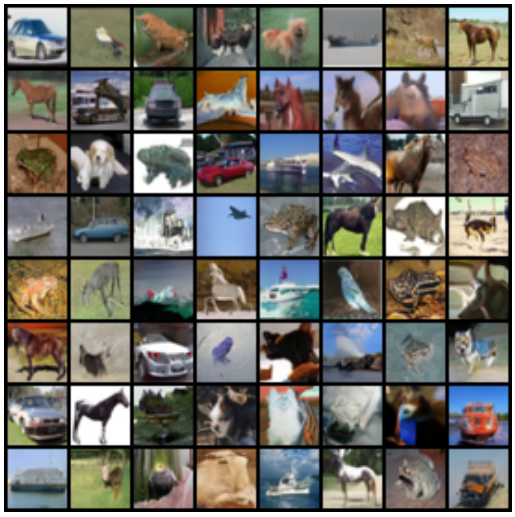


(b) $L_{2,1}$ FM

Figure 5: Generated samples for Patch noise with contamination level $\epsilon = 0.15$. The same source samples are used for squared and $L_{2,1}$ Flow Matching.

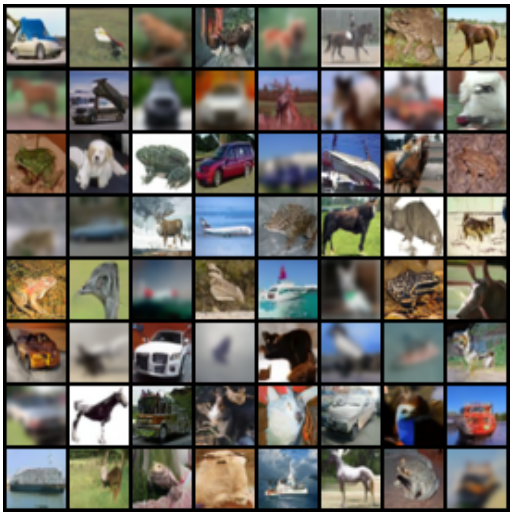


(a) Squared FM

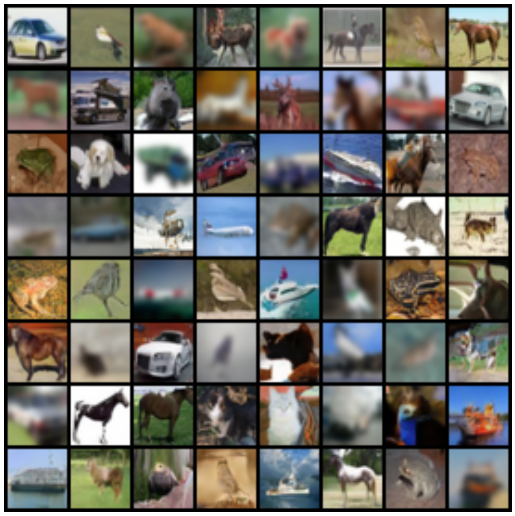


(b) $L_{2,1}$ FM

Figure 6: Generated samples for Inversion with contamination level $\epsilon = 0.15$. The same source samples are used for squared and $L_{2,1}$ Flow Matching.

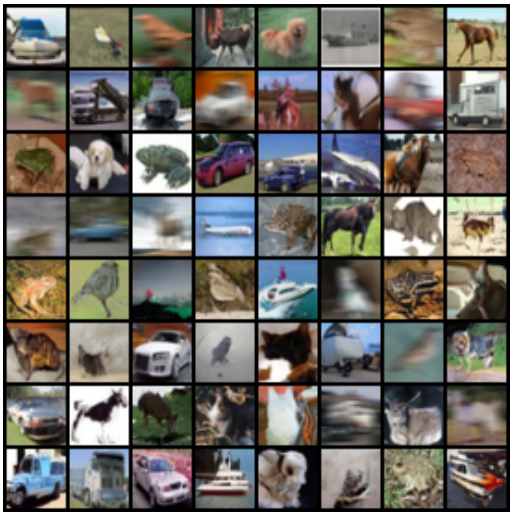


(a) Squared FM

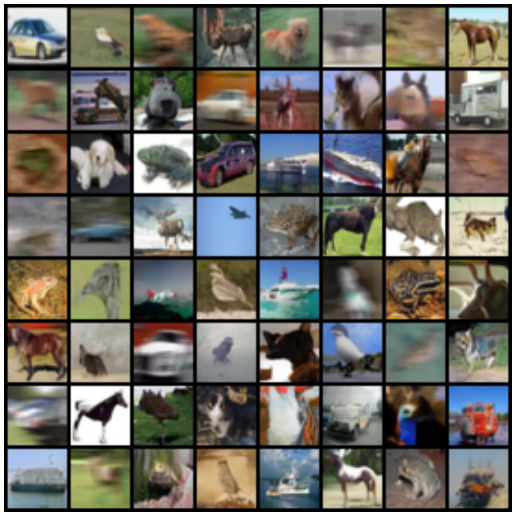


(b) $L_{2,1}$ FM

Figure 7: Generated samples for Gaussian blur with contamination level $\epsilon = 0.15$. The same source samples are used for squared and $L_{2,1}$ Flow Matching.



(a) Squared FM



(b) $L_{2,1}$ FM

Figure 8: Generated samples for Motion blur with contamination level $\epsilon = 0.15$. The same source samples are used for squared and $L_{2,1}$ Flow Matching.