COLLISION AVOIDANCE AND SHAPE FIDELITY IN FLOW-BASED SWARM FORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Collision avoidance and shape fidelity impose distinct requirements on flow-based swarm formation. We study their relationship using a frozen conditional flow-matching model. On airplane and car, learned second-order corrections with selected damping do not significantly outperform first-order controls. Across training seeds of potential-based corrections, models that produce more collision-free rollouts also place agents farther from the generated reference surface. Two loss interventions alter geometric deviation and collision performance without eliminating close contacts. We then construct a fixed pairwise correction from empirical distance statistics and integrate it with a displacement cap and adaptive sub-steps. Without retraining, it achieves 500 clean rollouts out of 500 across five agent counts up to 512 in each of five ShapeNet categories. Optimal reciprocal collision avoidance, applied within the sampling loop at its released settings, matches these collision counts, but the methods differ in their effects on formation geometry. Ablations show that integration controls affect collision performance and that similar collision counts can accompany substantially different geometric errors.

1 INTRODUCTION

Conditional flow-matching models generate point trajectories that can be used to construct swarm formations (Plou et al., 2024; Lipman et al., 2023). In this setting, each point represents an agent, and successful formation requires both adequate inter-agent separation and fidelity to the intended shape. Improving separation can compromise shape fidelity: dispersing agents may reduce collisions while moving them away from the target surface (Figure 1). We study these requirements jointly, examining how the dynamics, parameterization, and numerical integration of corrections applied during sampling affect collision avoidance and formation geometry.

Collision avoidance can be incorporated into a frozen generative model through reciprocal velocity adjustments, sampling-time constraints, or learned interactions between agents. Optimal reciprocal collision avoidance (ORCA) computes local velocity adjustments to avoid pairwise collisions (van den Berg et al., 2011), and Gen-Swarms applies it inside the sampling loop of the model we use (Plou et al., 2024); constrained flow samplers modify intermediate states to satisfy specified constraints (Utkarsh et al., 2025; Liang et al., 2025b). These approaches differ in how they modify the generated trajectories. We focus on learned corrections and a fixed pairwise correction, using the base model's in-loop ORCA as the collision-avoidance baseline.

We compare learned corrections that vary in dynamical order and parameterization: first-order velocity corrections or damped second-order acceleration corrections, represented by vector fields or potential gradients. All corrections use the same frozen base checkpoint within each category and the same evaluation seeds. This design preserves a common geometric reference while allowing us to compare collision performance across correction models.

On airplane and car, second-order corrections with damping selected by head type do not significantly outperform first-order controls; one performs close to the uncorrected model. Across training seeds of potential-based heads, models that produce more collision-free rollouts also place agents farther from the generated reference surface. Trajectory-wide collision penalties and stronger surface regularization alter geometric deviation and collision performance, but neither eliminates close contacts in the tested configurations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

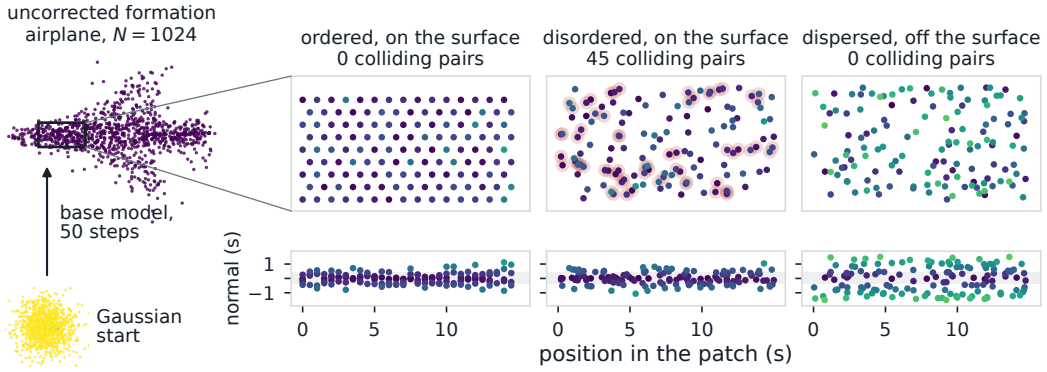


Figure 1: Left: the initial Gaussian cloud (evaluation seed 0; shown at 1/7.4 scale) and its uncorrected airplane formation at $N = 1024$, viewed from above; the box marks the patch size. Right: three constructed configurations of 135 agents on a patch of that size, from above (top) and from the side (bottom), at one scale in target spacings s. The ordered and dispersed configurations have no pair closer than $\kappa = 0.02$ model units; the disordered one has 45 (faint red halos). Color indicates surface distance, from purple (0) to yellow ($2s$ or more); the left panel uses the reference in Supplementary Figure S4. The band extends one reference thickness, $0.39s$, on either side. The constructed patches illustrate why collision counts alone do not determine formation geometry; they are not rollout results.

We then evaluate a fixed pairwise correction constructed from empirical pair-distance statistics, with a displacement cap and adaptive sub-steps. It achieves 500 clean rollouts out of 500 across the five evaluated agent counts in each of five ShapeNet categories, without retraining. ORCA at its released settings matches these counts, with geometric differences that depend on the category. Ablations identify substantial effects of the integration controls on collision performance and of the length-scale choice on distance from the reference surface.

2 Problem Setup and Evaluation

We use the conditional flow-matching model of Gen-Swarms (Plou et al., 2024), which transports a Gaussian cloud of N points into a ShapeNet formation (Chang et al., 2015) in 50 integration steps. We use its released checkpoints for airplane and car; for chair, bottle and table we trained one checkpoint each with the released training code and configuration for 60,000 iterations. All checkpoints remain frozen. We evaluate every correction family on airplane and car, and additionally evaluate the fixed correction and ORCA on chair, bottle, and table. In direct perturbation checks, changing the positions of other agents did not change an agent's base-model velocity. The corrections introduce interactions between agents during sampling.

We compare learned corrections, a fixed pairwise correction, and ORCA (van den Berg et al., 2011), using the uncorrected base model as a reference. Learned first-order heads add a velocity correction based on an agent's local neighborhood; second-order heads add an acceleration under fixed damping, $\gamma = 16$ for vector-field and $\gamma = 1$ for potential-based heads, each selected on separate selection seeds. We cross the order of the dynamics with two parameterizations, a vector field and the negative gradient of a scalar potential Ψ, and train eight independently initialized heads per configuration (Supplementary Section S1.1). Learned heads use 50 Euler steps for first-order dynamics and velocity Verlet for second-order dynamics, without the fixed correction's displacement cap or sub-steps. The fixed correction is defined in Section 5 and is evaluated on airplane and car using eight potentials tabulated from disjoint sets of base-model outputs. Uncertainty estimates over repeated learned models therefore describe variation across training seeds, while repeated fixed-correction evaluations describe variation across potential tabulations. ORCA is the base model's own layer, applied at every integration step with its released settings: agent radius 0.03, which requests a pair separation of 0.06 model units, neighbor distance 0.5, time horizon 0.05 and maximum speed 6.

We define a collision as a pairwise distance below $\kappa = 0.02$ model units at a scored integration step, steps 1 to 49 (Supplementary Section S2.2). A rollout is clean if no such violation is observed at any scored step. Distances are evaluated in the fixed model coordinate system without rescaling each generated cloud. The aggregate clean-rollout count sums the clean rollouts at $N \in \{32, 64, 128, 256, 512\}$, using 100 evaluation seeds at each agent count, for a maximum of 500. The same seeds are used across methods. We report results at $N = 1024$, over the same 100 evaluation seeds, separately from the aggregate clean-rollout count.

We measure geometric deviation at the final output (step 50) as each agent's distance to a dense uncorrected reference cloud for the same shape. Distances are normalized by the target spacing $s(N)$, defined as the category's median nearest-neighbor spacing at N agents and fitted as $s(N) = CN^{-P}$ for each category. We average the per-rollout means and 90th percentiles over evaluation seeds (distributions in Supplementary Figure S5). Because the reference is generated by the base model, this quantity measures proximity to that model's output. A small distance alone does not establish coverage of the reference or fidelity to the ShapeNet mesh.

The reference thickness, the median spread of target points along the local surface normal over 30 training clouds, is 0.27 target spacings at $N = 512$ on airplane and 0.33 on car. We use it as descriptive context rather than a criterion for declaring a rollout collision-free. Shape fidelity against held-out point clouds is evaluated separately using minimum matching distance, coverage, and 1-NN accuracy under Chamfer distance at $N = 1024$ and 2048 (Supplementary Table S6). We express crowding as $\kappa/s(N)$, which increases with agent count. Supplementary Table S4 reports the category comparisons.

3 Learned corrections

Second-order corrections have lower mean aggregate clean-rollout counts in all four comparisons with first-order controls (Table 1). For example, the airplane vector-field configuration decreases from 270 ± 23 to 245 ± 33 clean rollouts out of 500. Exact two-sided permutation tests on the eight per-run counts identify lower second-order means for airplane potential-based and car vector-field heads after Holm correction across the four comparisons ($p < 0.05$; Supplementary Table S1). Car second-order vector-field heads score 275 ± 6, compared with 253 for the uncorrected model, and show no improvement over it on selection seeds. Vector-field heads perform better than potential-based heads on airplane, whereas the reverse holds on car. Eight retrainings per configuration with independent training streams also find no significant second-order improvement, although the significant differences change (Supplementary Table S3).

A separate damping sweep provides a complementary comparison of the two dynamical orders (Supplementary Figure S2). The correction is rescaled by $2/\Delta t$, making the update at $\gamma\Delta t = 2$ algebraically identical to the first-order update. Second-order violations also span more of the integration horizon, while first-order violations concentrate in the last three steps (Supplementary Figure S3). These observations are consistent with a role for retained momentum, although differences in optimization and training difficulty remain possible explanations.

Figure 2 relates the aggregate clean-rollout count to mean reference-surface distance at $N = 512$. Across eight training seeds per configuration, the rank correlations are 0.81–0.98 for potential-based heads and 0.21–0.55 for vector-field heads. At the 5% level, these correlations are significant only for potential-based heads. The fixed correction reaches 500 clean rollouts with a mean reference-surface distance below half a target spacing at this agent count. Supplementary Figure S4 shows example endpoint clouds at $N = 1024$.

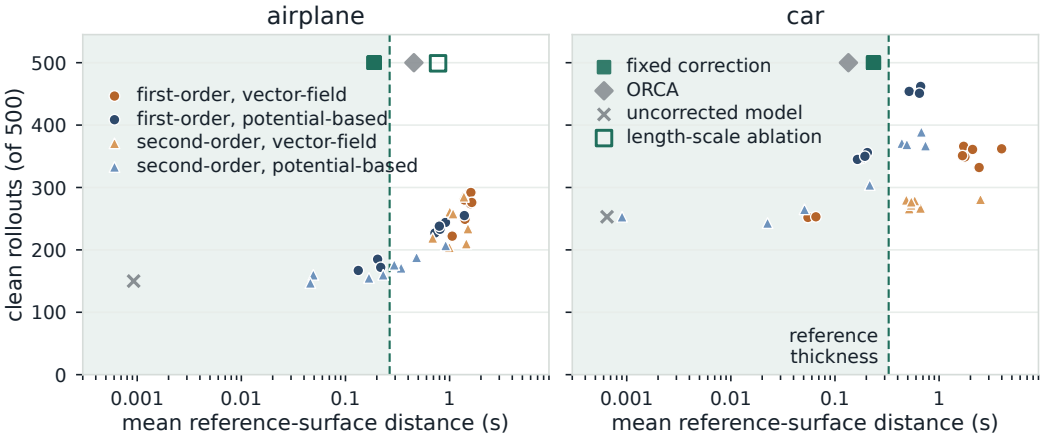


Figure 2: Aggregate clean-rollout count versus mean reference-surface distance at $N = 512$ (target spacings, logarithmic axis), one point per training seed for each learned configuration: circles first order, triangles second order, color the head type. The fixed correction, ORCA and the uncorrected model are shown for comparison; the open square is the length-scale ablation (Table 2), and the shaded band is each category's reference thickness at $N = 512$. Rank correlations are 0.81–0.98 for potential-based heads and 0.21–0.55 for vector-field heads.

4 Analysis of learned corrections

We measure local separation using $g(\kappa)$, the mean number of neighbors within the collision radius per agent, divided by a reference count extrapolated from a power-law fit over the first-neighbor shell (Supplementary Section S2.1). The statistic is 0 without contacts and 1 when observed and reference counts agree; by construction, it also equals 1 whenever the mean count reaches one. Figure 3 reports it from $N = 256$ to 2048, using five evaluation seeds per point, for one trained head per learned configuration and one tabulation for the fixed correction.

The lowest learned values range from 0.11 to 0.54 and approach the uncorrected model's 0.92–0.99 as crowding increases, partly because the statistic saturates when the mean neighbor count reaches one. Every learned configuration therefore has contacts within κ at every plotted setting. ORCA attains zero observed contact correlation at every plotted setting, and the fixed correction at all but one. Their spatial distributions nevertheless differ: on airplane, the fitted exponent of ORCA's neighbor count, an effective-dimension estimate, increases from 2.98 to 3.82 between 256 and 2048 agents, while the fixed correction's remains between 2.68 and 2.76 (Supplementary Table S5). Values above 3 arise because the fit begins at the edge of the exclusion region, where the count rises faster than a power law; ORCA's released radius places that edge at a separation of 3κ.

The separation term compares the sorted nearest-neighbor distances of the rollout endpoint with those of the target. Configurations with identical nearest-neighbor-distance distributions consequently receive the same value of this term, even if their spatial arrangements differ. The surface term, a one-sided Chamfer distance from agents to the target, provides an additional preference for proximity to the target. The dependence of the separation term on distance statistics motivates a possible explanation for the observed tradeoff, but does not establish a limitation of the complete training objective.

One geometric hypothesis is that displacement normal to a locally planar surface can increase an agent's distances from surrounding neighbors while moving it away from the target. At exact coplanarity, the first derivative of those distances in the normal direction is zero, so this observation alone does not establish a preferred gradient direction or explain which configurations training will select.

We also trained auxiliary potentials that depend only on pair separation, using the same objective, either from scratch or from a fit to the tabulated potential. At 0.3 target spacings, the scaled outward derivative with respect to normalized separation is 0.24 for the scratch-trained potentials and 0.25 for the implemented, smoothed potential (Section 5). The target histogram's first occupied bin is

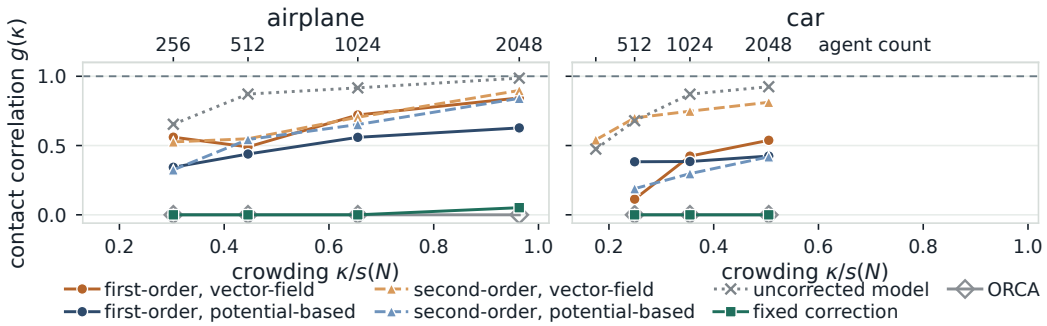


Figure 3: Contact correlation $g(\kappa)$ versus crowding $\kappa/s(N)$ at the last scored step, five evaluation seeds per point, for training seed 0 of each learned configuration and tabulation 0 of the fixed correction; the top axis gives N, and both panels share the crowding axis. Solid lines are first-order and dashed lines second-order learned configurations; dotted is the uncorrected model. A value of 1 is the configuration's own extrapolated reference and 0 no observed contacts; settings where that reference expects fewer than three contacting pairs per rollout are omitted. Only the fixed correction and ORCA reach zero, a finite-sample result; the learned heads are integrated without the fixed correction's displacement cap and sub-steps.

centered at 0.60 spacings, compared with a collision radius of 0.45 (Supplementary Figure S1). This target-data statistic does not establish which separations the model encounters during training; that requires a separate analysis of the training states.

We tested two changes to the training objective for first-order corrections on airplane (Figure 4). First, we added a collision hinge evaluated at every integration step and trained eight models per head type from different random initializations. At $N = 512$, $g(\kappa)$ ranges from 0.30 to 0.61 for vector-field heads and from 0.41 to 0.69 for potential-based heads, compared with 0.49 and 0.44 for the endpoint-trained heads of Figure 3. Mean reference-surface distance is 2.6 target spacings for vector-field heads on selection seeds, compared with 1.46 for endpoint-trained heads on evaluation seeds. Close contacts remain in both configurations.

Second, we increased the surface-loss weight by factors of ten and one hundred for the vector-field head, holding the separation-term weight at its endpoint-training value, with four training seeds per setting. Mean reference-surface distance at $N = 512$ decreases from 1.46 target spacings for the endpoint-trained heads to 0.66–0.78 and then 0.48–0.55 on selection seeds. At $N = 128$, clean-rollout counts fall from a baseline mean of 66.8 out of 100 to 53–60 and then 9–10. Contact correlation at $N = 512$ increases to 0.59–0.72 and then 0.86–0.96. Stronger surface regularization improves reference proximity but worsens collision performance over the tested weight range.

5 Method

5.1 Pairwise correction

We add a fixed pairwise correction to the frozen base velocity field. Let $\mathbf{v}_i$ denote the base velocity of agent i, r_{ij} the distance between agents i and j, and $\hat{\mathbf{r}}_{ij}$ the unit vector from j to i. With target spacing $s = s(N)$ and a tabulated potential U, the unbounded Euler proposal is

$$\mathbf{x}_i^{\text{prop}} = \mathbf{x}_i + (\mathbf{v}_i + \mathbf{F}_i)\Delta t, \qquad \mathbf{F}_i = -\frac{\lambda}{s} \sum_{j:\, r_{ij} < r_{\text{cut}} s} U'\Big(\frac{r_{ij}}{s}\Big)\, \hat{\mathbf{r}}_{ij}. \tag{1}$$

We use $\lambda = 0.01$ and an interaction cutoff $r_{\text{cut}} = 1.5$. The base flow has 50 steps of length $\Delta t = 0.02$. The proposal in Eq. 1 is subject to the displacement cap and adaptive subdivision described below.

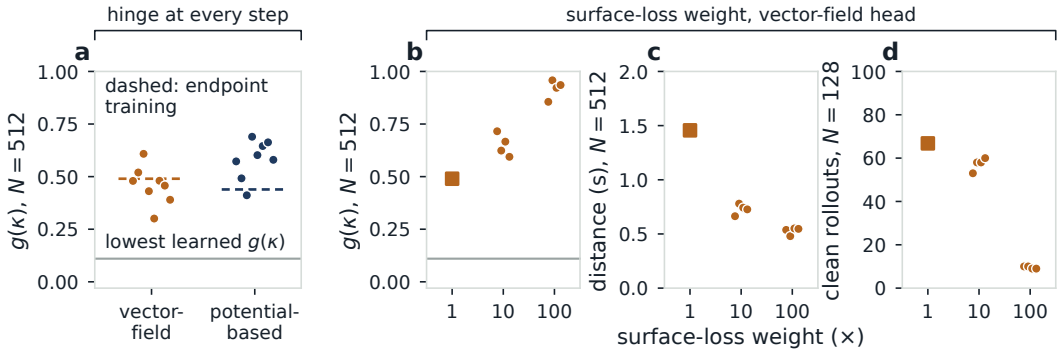


Figure 4: Two loss interventions on first-order corrections, airplane. (a) $g(\kappa)$ at $N = 512$ with the collision hinge applied at every step, eight runs per head; dashed lines mark endpoint training. (b–d) The vector-field head with its surface-loss weight raised 10- and 100-fold, four runs each; squares mark endpoint training, the head of Figure 3 in (b) and eight-run means in (c) and (d). Retrained heads are scored on selection seeds 200–299 and $g(\kappa)$ on evaluation seeds 0–4; the square in (c) uses evaluation seeds 0–99. The gray line in (a) and (b) is the lowest $g(\kappa)$ of any learned configuration.

5.2 Potential construction

For each category, we estimate pair-distance histograms from target clouds and uncorrected base-model outputs, using normalized separation $x = r/s$ and pooling over agent counts. Denoting these histograms by h_{target} and h_{base}, we initialize the potential with the empirical log-ratio

$$U(x) = -\ln \frac{h_{\text{target}}(x)}{h_{\text{base}}(x)}. \tag{2}$$

This construction is related to the distribution inversion used in effective pair-potential estimation (Reith et al., 2003). Here it provides a single estimate, without iterative refinement.

The histograms have 48 bins on $0 \le x < 2$, each normalized to unit mass, from the first 30 training-split clouds, farthest-point subsampled, and 25 base-model outputs at each of $N = 128$, 256 and 512. We set the potential to its cap of 5 below the smallest separation in the target histogram, clip negative values and truncate the interaction at 1.5 target spacings. We smooth U with a Gaussian kernel of standard deviation one bin, then differentiate and extend the steepest inner derivative toward zero separation (Supplementary Figure S7). The inward continuation supplies a nonzero correction within the region that would otherwise have a flat potential. Eight estimates that share the target clouds and use disjoint sets of base-model outputs are shown in Supplementary Figure S6; Supplementary Figure S1 compares the resulting pairwise correction with the learned force probes. No model parameters are trained for this construction.

5.3 Numerical integration

We limit the correction's contribution to each agent's displacement to $0.15s$ per sub-step, $\|\mathbf{F}_i\|\Delta t/k \le 0.15s$ for k sub-steps; the base velocity is not limited. We define x_{on} as the largest tabulated normalized separation below 1 where the smoothed U exceeds one tenth of its cap (0.854 on airplane). Each base-flow step uses $k = \min(\max(\lceil C \rceil, 1), 8)$ equal sub-steps. Here C is the largest positive pair-closing displacement predicted by base velocities within the cutoff, normalized by s and divided by the window $\max(x_{\text{on}} - \kappa/s,\, 0.1)$. The base velocity remains fixed during these sub-steps, while the pairwise correction is recomputed at each one. These controls limit correction displacements and adapt the integration step. Their effects on collision performance are evaluated separately in Section 6.

6 Results

The fixed correction achieves 500 clean rollouts out of 500 on airplane and car for each of eight potential tabulations, compared with 150 and 253 for the uncorrected base model (Table 1). ORCA

also achieves 500 in both categories. At $N = 1024$, the fixed correction averages 90.8 clean rollouts out of 100 on airplane and 100 on car; ORCA achieves 100 on both. At $N = 2048$, out of 50 rollouts, ORCA achieves 48 on airplane and 50 on car, compared with means of 0.6 and 48 for the fixed correction (Supplementary Table S10).

Without retuning the correction settings, the method also achieves 500 clean rollouts on chair, bottle, and table (Supplementary Table S4). To examine sensitivity to potential construction, we applied each category's potential to each category's base model. All 25 combinations are clean at the evaluated agent counts $N \in \{128, 256, 512\}$, but their reference-surface distances differ (Supplementary Figure S6). Thus, potential choice can affect geometry even when the evaluated collision counts agree.

	airplane				car			
method	Clean	1024	Ref.	ΔMMD	Clean	1024	Ref.	ΔMMD
uncorrected model	150	0	0.00	+0%	253	0	0.00	+0%
first order, vector field	270 ± 23	0.5	2.47	+18%	328 ± 48	1.5	2.60	+7%
first order, potential-based	215 ± 35	0.1	0.34	+193%	389 ± 55	15.6	0.32	+126%
second order, vector field	245 ± 33	0	2.06	+12%	275 ± 6	0	1.26	+5%
second order, potential-based	170 ± 20	0	**0.20**	+70%	320 ± 61	0.8	**0.23**	+43%
fixed correction	**500** ± 0	90.8	0.40	**−0%**	**500** ± 0	**100**	0.44	**−2%**
ORCA	**500**	**100**	0.87	+28%	**500**	**100**	0.40	+12%

Table 1: Collision and geometry results on airplane and car. Clean is the aggregate clean-rollout count out of 500; uncertainty is the standard deviation over eight training runs or potential tabulations. The $N = 1024$ column reports clean rollouts out of 100 over evaluation seeds 0–99, averaged over those runs or tabulations. Ref. is the 90th-percentile reference-surface distance at $N = 512$, in target spacings, averaged likewise. ΔMMD is the rounded percentage change in minimum matching distance at $N = 2048$ relative to the uncorrected model, with ORCA at its released radius. Bold marks the best value in each column among the corrections.

Across the five categories at $N = 512$, the fixed correction's 90th-percentile reference-surface distance ranges from 0.31 to 0.45 target spacings, compared with 0.25 to 0.87 for ORCA (Supplementary Table S4). The fixed correction has lower error on airplane and bottle; ORCA has lower error on car, chair, and table. Neither method uniformly dominates this metric.

Held-out shape metrics give a different comparison (Supplementary Table S6). On airplane and car at $N = 1024$ and 2048, the fixed correction's minimum matching distance is 0.4–2.7% lower than the base model's, and coverage remains within 1.5 percentage points. ORCA at its released radius increases minimum matching distance on airplane by approximately 10% and 28% at the two agent counts. On car, it decreases this distance by approximately 1% at $N = 1024$ but increases it by 12% at $N = 2048$, where coverage falls from 29.0% to 17.5%. The fixed correction also changes some distributional statistics substantially: car 1-NN accuracy increases from 57.7% to 75.7% at $N = 1024$. Its near-baseline matching distance and coverage therefore do not establish that the generated distribution is unchanged.

We evaluated the correction's components through individual removals or replacements on airplane (Table 2). Removing the displacement cap reduced the aggregate clean-rollout count from 500 to 341, while disabling adaptive subdivision reduced it to 460. Removing the inward continuation reduced the count to 467. Replacing the tabulated potential with a linear ramp preserved the evaluated collision counts, including 24 of 25 clean rollouts at $N = 1024$; this replacement changes the potential-height threshold and the ramp width that sets adaptive subdivision, as well as the potential's shape. Retaining the attractive branch also preserved these counts. These results show a strong dependence on the integration controls, while the collision metrics did not distinguish the tabulated potential from its linear replacement.

Holding the correction's scale fixed at $s(128)$ across agent counts changes the aggregate clean-rollout count only from 500 to 499 and the $N = 1024$ count from 24 to 23 out of 25, while increasing the 90th-percentile reference-surface distance at $N = 512$ from 0.39 to 1.24 target spacings (Table 2). Nearly identical collision counts can therefore accompany substantially different geo-

variant	Clean	1024 (of 25)	Ref.
fixed correction	500	24	0.39
integration			
no displacement cap	**341**	**0**	**0.70**
no adaptive subdivision ($k = 1$)	**460**	**4**	0.38
potential			
no inward continuation (flat plateau)	**467**	23	0.39
scale fixed at $s(128)$ for every N	499	23	**1.24**
linear ramp in place of the tabulated shape	500	24	0.32
attractive branch kept (not clipped)	500	24	**0.46**

Table 2: Component ablations on airplane. Clean is the aggregate count out of 500, and the 1024 column gives clean rollouts out of 25 at $N = 1024$ (evaluation seeds 0–24). Ref. is the 90th-percentile reference-surface distance at $N = 512$, in target spacings, measured in separate ablation rollouts on evaluation seeds 0–24, so the fixed correction's value differs slightly from Table 1. The fixed-scale variant uses $s(128)$ for the force, the cutoff, the displacement cap and the sub-step window at every N. Each modification is applied separately, so changes are not additive. Bold marks values worse than the fixed correction's by more than 2 rollouts or 0.05 target spacings.

metric deviations. Preliminary shared-GPU timings indicate that the cell-list implementation adds approximately 4–7 milliseconds per base-flow step for $N = 128$ to 8192 (Supplementary Table S7).

The fixed correction has nonzero observed contact correlation on airplane at $N = 2048$, where $g(\kappa) = 0.052$. The other 19 category–agent-count settings have zero observed contact correlation. Supplementary Figure S8 relates these observations to crowding, expressed as the collision radius divided by the target spacing, relative to each category's potential-height threshold. This comparison motivates an empirical operating-range analysis, but does not establish a universal capacity bound.

In a separate long-horizon experiment on airplane at $N = 1024$, over evaluation seeds 0–99, we continue each rollout to step 250. When the base velocity is set to zero after the 50 base-flow steps, so that only the correction moves the agents, no rollout has a violation from step 52 to step 250, and the median nearest-neighbor distance settles at $1.11s$. Continuing the base flow at its final time produces violations after about step 60: 84 rollouts are violation-free at step 100 and 39 at step 250, but only 2 remain so throughout steps 52–250. The uncorrected model's median nearest-neighbor distance falls to $0.02s$ (Supplementary Figure S9).

7 Related work

Evaluation of point-cloud formations. Held-out shape quality is usually measured by minimum matching distance and coverage (Achlioptas et al., 2018) and by 1-NN accuracy (Yang et al., 2019), whose Chamfer-based forms have known limitations (Bastico et al., 2026); Gen-Swarms evaluates generated shapes alongside collision and trajectory metrics (Plou et al., 2024). Other domains also evaluate physical validity or inter-object relations: clashes in docking poses (Buttenschoen et al., 2024), macro overlap in chip placement (Xie et al., 2026), physical errors of learned particle simulators (Toshev et al., 2023) and clearance between tree crowns (Yang & Liu, 2026). We evaluate clean-rollout counts, distance to a generated reference surface, and local pair statistics within one correction framework.

Formation display. Pairing goal generation with reciprocal collision avoidance predates generative models: multi-robot displays assign image-derived goals and move robots with ORCA (Alonso-Mora et al., 2012), and drone light-show planning assigns drones and plans minimum-snap trajectories under pairwise separation constraints (Alqudsi, 2026). Concurrent work generates light shows with video models, keeps tracked image points apart with a soft repulsion and enforces separation with a hard-constraint safety filter (Reinhold et al., 2026).

Constraints during generative sampling. Several methods impose constraints on a pretrained model at inference time. PCFM projects intermediate states onto nonlinear equality con-

straints (Utkarsh et al., 2025), ECI uses a gradient-free extrapolate–correct–interpolate procedure (Cheng et al., 2025), HardFlow constrains the final sample by trajectory optimization (Li et al., 2026b), and chance-constrained flow matching accounts for the distortion that repeated projection causes (Liang et al., 2025b). For multiple robots, projected diffusion enforces inter-robot separation at every sampling step, demonstrated for up to nine robots at tens of seconds to minutes per plan (Liang et al., 2025a), and collision-constrained flow matching targets selected adversarial collisions (Li et al., 2026a). Constraint steps also enforce separation or non-overlap among many objects elsewhere: learned neural projections handle collisions in point-based simulation (Yang et al., 2020), discrete-guided diffusion repairs trajectories for up to 100 robots (Liang et al., 2026), and FlowPlace samples overlap-free chip layouts (Xie et al., 2026). Enforcing separation among up to 2048 agents entails many pairwise constraints; ORCA provides a direct baseline using reciprocal velocity adjustments (van den Berg et al., 2011).

Pair potentials and second-order dynamics. The potential construction is related to Boltzmann inversion of pair-distance distributions (Reith et al., 2003), although we use a single empirical estimate rather than an iterative procedure; pairwise repulsion also appears as a training loss in point-cloud upsampling (Yu et al., 2018). Damped second-order population models (Guan et al., 2026) and flow-matched motion fields that include accelerations (Nguyen et al., 2025) motivate acceleration-based corrections. Our contribution is the empirical comparison of collision avoidance and geometry, including the effects of correction design and integration.

8 Discussion

The evaluated corrections expose a tradeoff between collision performance and formation geometry. Learned potential-based models that produce more collision-free rollouts also tend to place agents farther from the reference surface. Changing the collision-loss evaluation or surface-loss weight alters this relationship without eliminating close contacts in the tested settings. The dependence of the separation loss on distance statistics offers a possible explanation, but the experiments do not show that the complete objective precludes a separated, surface-preserving solution.

Without retraining or retuning, the fixed correction achieves 500 clean rollouts in 16 of 21 categories. Four categories fall short, and all faucet potentials are rejected by the implementation (Supplementary Table S11 and Figure S8). Its comparison with ORCA depends on the category and geometric metric, and its ablations show that numerical integration affects collision counts while the length-scale choice can substantially change reference-surface distance. A useful extension would test objectives that explicitly control displacement normal to the reference surface.

The scope of our conclusions is limited to one base-model family, two categories for learned corrections, and, for the fixed correction, five categories in the main evaluation and 16 additional categories in the supplement. Reference-surface distance measures proximity to a generated cloud and cannot establish mesh fidelity or coverage by itself. Collisions are checked at scored integration steps. Checking straight segments between steps preserves all collision-free classifications for the fixed correction at $N = 32$ to 512, but reduces the mean clean-rollout count by 5 out of 100 on airplane at $N = 1024$ (Supplementary Table S10); neither check establishes separation for physical vehicles. Within this scope, the results support evaluating collision avoidance together with reference geometry and held-out shape quality, including larger populations where collision-free performance deteriorates.

AI use statement

In this work, we used generative AI tools, Claude models (Anthropic) operated through the Claude Code agent, to refine hypotheses, to implement the corrections and the training and evaluation code, and to interpret results. We have not used generative AI tools to develop the analysis of Section 4 or formulate its mathematical statements, to generate data sets, or for translation, and the tasks involving proofs, surveys, interviews or qualitative data analysis are not applicable to this work. Additionally, we used these tools to write and edit software, including the scripts that subsample and reformat the ShapeNet point clouds, suggest experimental parameters, create the figures and tables, search for and summarize related literature, draft and edit the text, and format references. We have reviewed all AI-assisted work: the authors verified and tested all code, and all code and generated artifacts

were reviewed independently. In addition, hypotheses and their predicted outcomes were recorded before the corresponding experiments were run, force implementations were checked on two-agent and zero-force configurations, the figures and tables are regenerated from stored evaluation records by scripts, and every cited reference was checked against external bibliographic records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work studies collision avoidance in simulated swarm formation and involves no human subjects or personal data. The target shapes come from ShapeNet (Chang et al., 2015), used under its terms for non-commercial research. Formation control for drone swarms has civilian applications, such as aerial displays and inspection, and also possible surveillance and military uses; this work does not target a specific application.

Reproducibility statement

Section 2 specifies the base model, categories, evaluation seeds, collision definition and metrics, and Section 5 gives the correction, the potential construction and the integration settings. The airplane and car base checkpoints are the publicly released ones (Plou et al., 2024), whose code is released under the AGPL-3.0 license; the chair, bottle and table checkpoints were trained with that code and its released configuration for 60,000 iterations, and the sofa, bench, rifle and lamp checkpoints of Supplementary Table S11 with that code, a Gaussian prior and 200,000 iterations. The potentials are tabulated from, and the learned heads trained on, the ShapeNet training split; the held-out shape metrics use the test split (Supplementary Table S6). All methods are evaluated on the same seeds, 0–99 at each agent count (0–49 at $N = 2048$ and 0–24 for the ablations' $N = 1024$ column), except the held-out shape metrics, which use 200 generated clouds per method from seeds 998–1197; Supplementary Table S2 reports every training seed and potential tabulation. The loss interventions, damping sweep, integrator comparisons, cross-category transfer, staggered-release experiments and the ramp's spacing comparison use separate selection seeds (Figure 4; Supplementary Tables S8, S9 and S10; Supplementary Figures S2 and S6), and the long-horizon experiment continues the evaluation rollouts beyond step 50 (Supplementary Figure S9). Our code, which contains all experimental settings and values needed to reproduce the results as presented, will be made publicly available.

References

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning Representations and Generative Models for 3D Point Clouds. In *ICML 2018 (PMLR 80; arXiv:1707.02392)*, 2018. doi: 10.48550/arXiv.1707.02392.

Javier Alonso-Mora, Andreas Breitenmoser, Martin Rufli, Roland Siegwart, and Paul Beardsley. Image and animation display with multiple mobile robots. *The International Journal of Robotics Research*, 31(6):753–773, 2012. doi: 10.1177/0278364912442095.

Yunes Alqudsi. Enhancing Drone Light Shows Performances: Optimal Allocation and Trajectories for Swarm Drone Formations. *Journal of Field Robotics*, 2026. doi: 10.1002/rob.70318.

Matteo Bastico, David Ryckelynck, Laurent Corté, Yannick Tillier, and Etienne Decencière. Rethinking Metrics and Diffusion Architecture for 3D Point Cloud Generation. In *International Conference on 3D Vision (3DV)*, 2026. doi: 10.1109/3DV69130.2026.00133.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024. doi: 10.1039/d3sc04185a.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. arXiv:1512.03012, 2015.

Chaoran Cheng, Boran Han, Danielle C. Maddix, Abdul Fatir Ansari, Andrew Stuart, Michael W. Mahoney, and Yuyang Wang. Gradient-Free Generation for Hard-Constrained Systems. In *ICLR 2025 (arXiv:2412.01786)*, 2025. doi: 10.48550/arXiv.2412.01786.

Vincent Guan, Lazar Atanackovic, and Kirill Neklyudov. A Call to Lagrangian Action: Learning Population Mechanics from Temporal Snapshots. In *ICML 2026 (arXiv:2605.08550)*, 2026. doi: 10.48550/arXiv.2605.08550.

Ke Li, Kaidi Liang, Yuxin Ding, Debojyoti Biswas, Xianbiao Hu, and Ruwen Qin. CCFM: Collision-Constrained Flow Matching for Safety-Critical Scenario Generation. In *ECCV 2026 (arXiv:2607.04451)*, 2026a. doi: 10.1007/978-3-032-37490-5_5.

Zeyang Li, Kaveh Alim, and Navid Azizan. HardFlow: Hard-Constrained Sampling for Flow-Matching Models via Trajectory Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):13619–13635, 2026b. doi: 10.1109/TPAMI.2026.3708824.

Jinhao Liang, Jacob K. Christopher, Sven Koenig, and Ferdinando Fioretto. Simultaneous Multi-Robot Motion Planning with Projected Diffusion Models. In *ICML 2025 (PMLR 267; arXiv:2502.03607)*, 2025a. doi: 10.48550/arXiv.2502.03607.

Jinhao Liang, Yixuan Sun, Anirban Samaddar, Sandeep Madireddy, and Ferdinando Fioretto. Chance-constrained Flow Matching for High-Fidelity Constraint-aware Generation. arXiv:2509.25157, 2025b.

Jinhao Liang, Sven Koenig, and Ferdinando Fioretto. Discrete-guided diffusion for scalable and safe multi-robot motion planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23417–23424, 2026. doi: 10.1609/aaai.v40i28.39512.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *ICLR 2023*, 2023. doi: 10.48550/arXiv.2210.02747.

Khang Nguyen, An T. Le, Tien Pham, Manfred Huber, Jan Peters, and Minh Nhat Vu. FlowMP: Learning Motion Fields for Robot Planning with Conditional Flow Matching. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025. doi: 10.1109/IROS60139.2025.11246537.

Carlos Plou, Pablo Pueyo, Ruben Martinez-Cantin, Mac Schwager, Ana C. Murillo, and Eduardo Montijano. Gen-Swarms: Adapting Deep Generative Models to Swarms of Drones. In *ECCV 2024 Workshops (arXiv:2408.15899)*, 2024. doi: 10.1007/978-3-031-91813-1_6.

Pascal Reinhold, Alexander Gräfe, and Sebastian Trimpe. Generative AI for Safe and Photorealistic Drone Light Shows. arXiv:2606.25458, 2026.

Dirk Reith, Mathias Pütz, and Florian Müller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *Journal of Computational Chemistry*, 24(13):1624–1636, 2003. doi: 10.1002/jcc.10307.

Artur P. Toshev, Gianluca Galletti, Fabian Fritz, Stefan Adami, and Nikolaus A. Adams. LagrangeBench: A Lagrangian fluid mechanics benchmarking suite. In *Advances in Neural Information Processing Systems (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. doi: 10.52202/075280-2830.

Utkarsh Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gomez-Bombarelli, and Christopher Rackauckas. Physics-Constrained Flow Matching: Sampling Generative Models with Hard Constraints. In *NeurIPS 2025 (arXiv:2506.04171)*, 2025. doi: 10.52202/085713-5354.

Jur van den Berg, Stephen J. Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-Body Collision Avoidance. In *International Symposium of Robotics Research*, 2011. doi: 10.1007/978-3-642-19457-3_1.

Peng Xie, Ke Xue, Yunqi Shi, Ruo-Tong Chen, Chengrui Gao, Siyuan Xu, Chenjian Ding, Mingxuan Yuan, and Chao Qian. FlowPlace: Flow matching for chip placement. In *Design Automation Conference (DAC 2026; arXiv:2604.23658)*, 2026.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. PointFlow: 3D Point Cloud Generation with Continuous Normalizing Flows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. doi: 10.1109/ICCV.2019.00464.

Guang Yang and Fengchen Liu. Growing a stand, not a tree: Joint canopy generation reproduces crown shyness. *arXiv preprint arXiv:2608.28692*, 2026.

Shuqi Yang, Xingzhe He, and Bo Zhu. Learning physical constraints with neural projections. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. PU-Net: Point Cloud Upsampling Network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00295.