ConTiCoM-3D: A Continuous-Time Consistency Model for 3D Point Cloud Generation

Supplementary Material

7. Training

7.1. Architecture

We instantiate F_θ with a PVCNN-style UNet. The network consumes a point set $x_t \in \mathbb{R}^{B\times 3\times N}$ and a scalar timestamp t, and predicts a velocity field $F_\theta(x_t/\sigma_d, t) \in \mathbb{R}^{B\times 3\times N}$. Time is injected through a sinusoidal embedding (such as in positional encodings [46]) followed by a two-layer MLP, which is added to the global features of the bottleneck.

The encoder consists of four stages, each implemented as a *PVConv* block that fuses voxelized 3D convolutional features with pointwise MLP features, following PVCNN [25]. The channel schedule follows

$$\text{dims} = [64,\ 128,\ 256,\ 512,\ 1024],$$

corresponding to `dim_mults`=[2,4,8,16]. At stage ℓ, the voxel resolution is set to $r_\ell = 8 \cdot 2^{\ell-1}$. The bottleneck aggregates features with global max pooling, adds the time embedding, and refines the representation with two SharedMLP+LayerNorm blocks.

The decoder mirrors the encoder through four *PointNet++ Feature Propagation (FP)* modules [35]. Each FP block interpolates features from a coarser resolution back to a finer one, concatenates them with the corresponding encoder skip features, and processes the result with a SharedMLP. Finally, a SharedMLP head outputs three channels, which correspond to the predicted velocity vectors.

The complete layer structure is summarized in Table 5, which reflects the implementation with `dim_mults`=[2,4,8,16].

7.2. Hyperparameter Settings

We follow the TrigFlow parameterization [27], also given in the main paper (Eq. 9), with $\sigma_d = 1.0$ and $T_{\max} = \frac{\pi}{2}$:

$$x_t = \cos t\, x_0 + \sin(t)\, z, \quad z \sim \mathcal{N}(0, \sigma_d^2 I).$$

Objective. Training is based on single-time supervision combining analytic Flow Matching [24] (FM) and Chamfer reconstruction [9]. The FM loss is identical to Eq. 12 in the main paper. For completeness, we restate the Chamfer reconstruction loss (Eq. 13):

$$\mathcal{L}_{\text{CD}} = \text{CD}(f_\theta(x_t, t), x_0),$$

where CD denotes the *symmetric squared Chamfer distance*, normalized by the number of points. The total training objective (Eq. 14) is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{CD}}(t)\, \mathcal{L}_{\text{CD}},$$

with $\lambda_{\text{CD}}(t)$ a deterministic time-dependent weight, consistent with the main paper.

Training. We train using Adam [20] with learning rate 1×10^{-4} and gradient clipping at 1.0. The batch size is 64, and the schedule runs for 300 epochs. Point clouds are normalized to zero mean and unit radius, with 2048 points per shape.

Chamfer weighting. We use the deterministic schedule from the main paper:

$$\lambda_{\text{CD}}(t) = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \cos^2(t).$$

We set $\lambda_{\min} = 0.1, \lambda_{\max} = 0.3$, and $t \in [0, T_{\max}]$ with $T_{\max} = \pi/2$. This downweights CD at high noise and emphasizes it near the data manifold.

Note. As shown in Proposition 1 and proved in Appendix 8, minimization to zero of the FM loss implies $f_\theta(x_t, t) = x_0$ for all $t \in [0, T_{\max}]$. This means that a perfect velocity regression suffices for exact reconstruction.

Sampling. Single-step generation uses the predictor defined in main paper Eq. 16:

$$\hat{x}_0 = f_\theta(x_T, T_{\max}), \quad x_T \sim \mathcal{N}(0, \sigma_d^2 I),$$

achieving close to real-time inference. For optional refinement, we integrate the reverse PF-ODE (Eq. 17) using S steps:

$$x_{t-\Delta} = x_t - \Delta\, \sigma_d\, F_\theta(x_t/\sigma_d, t), \quad \Delta = T_{\max}/S.$$

By default, we use explicit Euler steps. Optionally, a *Heun proposal* [27] can be used for higher-order correction and local error control. Values $S = 2$ or $S = 4$ already improve fine-scale detail with minimal cost, while larger values of S provide diminishing returns.

8. Proof of Proposition

We restate Proposition 1 and provide a detailed derivation. We also discuss the role of the flow-matching loss and what happens if it is only approximately satisfied, linking this directly to reconstruction and generation quality.

Stage	Operation	Channels (in→out)	Notes
Input	–	3	Point coordinates
Time	Sinusoidal + MLP	$64 \to 1024$	Added at bottleneck
Enc-1	PVConv($k = 3$, res=8)	$3 \to 128$	$C_1 = 64 \cdot 2$
Enc-2	PVConv($k = 3$, res=16)	$128 \to 256$	$C_2 = 64 \cdot 4$
Enc-3	PVConv($k = 3$, res=32)	$256 \to 512$	$C_3 = 64 \cdot 8$
Enc-4	PVConv($k = 3$, res=64)	$512 \to 1024$	$C_4 = 64 \cdot 16$
Bottleneck	SharedMLP+LN $\times 2$	$1024 \to 1024$	add time feature
Dec-4	FP + skip(Enc-4)	$(1024 + 1024) \to 512$	upsample coords
Dec-3	FP + skip(Enc-3)	$(512 + 512) \to 256$	
Dec-2	FP + skip(Enc-2)	$(256 + 256) \to 128$	
Dec-1	FP + skip(Enc-1)	$(128 + 128) \to 64$	
Head	SharedMLP	$64 \to 3$	velocity channels

Table 5. ConTiCoM-3D architecture used for F_θ, instantiated with `dim_mults`=[2,4,8,16]. Encoder blocks are PVConv layers with voxel resolutions $[8, 16, 32, 64]$, the bottleneck injects time conditioning, and the decoder consists of FP upsampling modules with skip connections.

Proposition 1 (Closed-Form Recovery). Let the TrigFlow trajectory be given by Eq. 9:

$$x_t = \cos(t)\, x_0 + \sin(t)\, z, \qquad t \in [0, T_{\max}], \quad T_{\max} = \tfrac{\pi}{2},$$

where $x_0 \in \mathbb{R}^d$ is a data sample and $z \sim \mathcal{N}(0, \sigma_d^2 I)$. Define the predictor as in Eq. 10:

$$f_\theta(x_t, t) = \cos(t)\, x_t - \sin(t)\, \sigma_d F_\theta(x_t/\sigma_d, t).$$

The forward dynamics satisfy Eq. 11:

$$\frac{dx_t}{dt} = -\sin(t)\, x_0 + \cos(t)\, z.$$

The flow-matching loss is defined in Eq. 12:

$$\mathcal{L}_{\text{FM}} = \big\| \sigma_d F_\theta(x_t/\sigma_d, t) - (\cos(t)\, z - \sin(t)\, x_0) \big\|_2^2.$$

In the ideal zero-loss case, this implies the analytic regression target

$$\sigma_d F_\theta(x_t/\sigma_d, t) = \cos(t)\, z - \sin(t)\, x_0.$$

Then, for all $t \in [0, T_{\max}]$,

$$f_\theta(x_t, t) = x_0.$$

Proof. From Eq. 9,

$$x_t = \cos(t)\, x_0 + \sin(t)\, z.$$

Substituting into Eq. 10 gives

$$f_\theta(x_t, t) = \cos(t)\, x_t - \sin(t)\, \sigma_d F_\theta(x_t/\sigma_d, t).$$

Using the zero-loss condition implied by Eq. 12,

$$\sigma_d F_\theta(x_t/\sigma_d, t) = \cos(t)\, z - \sin(t)\, x_0,$$

we obtain

$$\begin{aligned} f_\theta(x_t, t) &= \cos(t)\big(\cos(t)\, x_0 + \sin(t)\, z\big) \\ &\quad - \sin(t)\big(\cos t\, z - \sin(t)\, x_0\big) \\ &= \cos^2(t)\, x_0 + \cos(t)\sin(t)\, z \\ &\quad - \sin(t)\cos(t)\, z + \sin^2(t)\, x_0 \\ &= (\cos^2(t) + \sin^2(t))\, x_0 \\ &= x_0. \end{aligned}$$

Thus, for every $t \in [0, T_{\max}]$, the predictor exactly recovers the ground-truth data point:

$$f_\theta(x_t, t) = x_0. \qquad \square$$

Discussion of Assumptions. The exact recovery critically depends on the flow-matching loss of Eq. 12. This objective enforces

$$\sigma_d F_\theta(x_t/\sigma_d, t) \approx \cos(t)\, z - \sin(t)\, x_0.$$

- In the **ideal case** (zero loss), the equality holds exactly.

- In **practice**, training and model capacity limitations imply

$$\sigma_d F_\theta(x_t/\sigma_d, t) = \cos(t)\, z - \sin(t)\, x_0 + \varepsilon_t,$$

where ε_t denotes the residual regression error.

Substituting this into Eq. 10 yields

$$f_\theta(x_t, t) = x_0 - \sin(t)\, \varepsilon_t.$$

Table 6. Loss ablation on ShapeNet [4] *airplane*. Metrics: ↓ lower is better, ↑ higher is better.

Variant	1-NNA (CD)↓	1-NNA (EMD)↓	MMD (CD)↓	COV↑	JSD↓
FM only	72.8	70.9	2.51	42.3	0.073
Chamfer only	83.5	62.0	3.84	19.7	0.109
FM + Chamfer ($\lambda_{\text{CD}}(t)$)	**69.9**	**64.8**	**2.18**	**51.6**	**0.061**

Consequences of Imperfect Recovery. There are three consequences of imperfect recovery:

Error amplification with time. The deviation from x_0 scales with $\sin(t)$. At small t, the error contribution is minimal, so the reconstructions remain close to exact. Near $T_{\max} = \pi/2$, however, the factor $\sin(t)$ is large and inaccuracies become most pronounced.

Impact on reconstruction quality. Imperfect cancellation of the noise z leads to visible artifacts: reconstructions may appear *blurry, distorted, or biased*. The model fails to perfectly invert Eq. 9, so residual noise contaminates the recovered sample.

Impact on generation quality. In generative sampling, these errors manifest as *loss of sharpness*, *inconsistent details*, or *mode bias* where the outputs drift toward regions favored by the imperfect predictor. The neat cancellation between cosine and sine terms, guaranteed by the zero-loss condition of Eq. 12, is thus the mechanism that allows clean recovery. When it fails, quality degradation becomes noticeable.

Conclusion. The derivation confirms that Proposition 1 holds rigorously under the ideal assumption implied by Eq. 12, with f_θ acting as an exact inverse mapping along the TrigFlow trajectory in Eq. 9. When the assumption is only approximately met, Eq. 10 yields systematic deviations amplified by $\sin(t)$, which explains the observed degradation in both reconstruction and generation quality. Proposition 1 therefore serves both as a theoretical guarantee and as a diagnostic principle linking analytic flow-matching to practical performance.

Remark. Proposition 1 highlights a precise correspondence between the flow-matching loss (Eq. 12) and perfect data recovery: when the loss is minimized to zero, reconstruction is exact; when it is nonzero, the resulting artifacts directly explain the observable drop in reconstruction and generation fidelity (Fig. 2 in the main text for empirical evidence).

9. Limitations of Teacher-Based Supervision in Point Cloud Models

A common strategy in consistency models (CMs) is **teacher–student distillation**, where a target function f_θ^- (often a stop-gradient or exponential moving average (EMA) of the current model) supervises predictions over time. The canonical distillation loss is introduced in previous CM works [17, 28, 45]:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{x_t,t}\left[\left\| f_\theta(x_t,t) - f_\theta^-(x_{t-\Delta t}, t-\Delta t)\right\|^2\right], \tag{18}$$

where $x_{t-\Delta t}$ is computed using a known generative path (e.g., a reverse ODE), and f_θ^- is a stop-gradient copy or an EMA teacher.

Although this has proven effective in image domains [28, 45], the approach fails when extended to 3D point cloud generation for several fundamental reasons:

(1) Absence of reliable perceptual metrics. In 2D vision, perceptual metrics such as LPIPS [51] or CLIP feature distances [36] align well with human similarity judgments. However, point clouds are unordered and sparse representations in $\mathbb{R}^3$ without canonical structure or correspondence. The lack of robust semantic metrics makes it hard to quantify the agreement between teacher and student outputs. Even if both $f_\theta(x_t,t)$ and $f_\theta^-(x_{t-\Delta t}, t-\Delta t)$ represent plausible reconstructions, the Euclidean distance in Eq. 18 may penalize them heavily due to point permutations or geometric shifts. In practice, Chamfer distance [9] and Earth Mover's Distance (EMD) [37] are the only metrics widely used for point clouds, and even they are limited in semantic sensitivity [1].

(2) Permutation and alignment mismatch. Let $X = \{x_i\}_{i=1}^M$ and $Y = \{y_j\}_{j=1}^M$ be predicted point clouds. Chamfer or EMD distances allow for set-wise comparison, but teacher–student losses like Eq. 18 assume fixed point correspondences. Mathematically, there exists a permutation π such that:

$$\min_{\pi \in S_M} \sum_{i=1}^{M} \left\| f_\theta^{(i)} - f_\theta^{-(\pi(i))} \right\|^2, \tag{19}$$

where S_M is the symmetric group. This matching is non-differentiable and thus incompatible with backpropagation. As a result, point-level losses are ill-posed for unordered sets, motivating permutation-invariant architectures such as PointNet and PointNet++ [34, 35].

Table 7. Effect of sampling steps (S) on ShapeNet [4] *car*.

Variant	1-NNA (CD)↓	EMD↓	Time (s)↓	Memory (GB)↓
$S = 1$ (Euler)	53.9	51.9	**0.22**	**5.1**
$S = 2$ (Euler)	**53.3**	**51.4**	0.41	5.2
$S = 4$ (Euler)	53.0	51.1	0.79	5.3
$S = 1$ (Heun)	53.5	51.7	0.25	5.2

(3) Inconsistent signal geometry. During early training, the teacher f_θ^- is itself unreliable and may regress to blurred or collapsed outputs. Since x_t and $x_{t-\Delta t}$ come from different noise levels, the corresponding targets $f_\theta^-(x_{t-\Delta t}, t - \Delta t)$ may lie on significantly different manifolds. The student is then penalized for following a different (yet plausible) geometric reconstruction. This causes instabilities and hurts convergence, consistent with reports of teacher collapse in distillation frameworks [5, 8, 39].

(4) Temporal drift and gradient mismatch. The underlying assumption of distillation is temporal smoothness [17, 45]:

$$f_\theta^-(x_{t-\Delta t}, t - \Delta t) \approx f_\theta^-(x_t, t) - \Delta t \cdot \partial_t f_\theta^-(x_t, t).$$

However, in practice, the numerical estimate

$$\frac{f_\theta^-(x_t, t) - f_\theta^-(x_{t-\Delta t}, t - \Delta t)}{\Delta t} \tag{20}$$

does not approximate $\partial_t f_\theta^-$ accurately for point clouds due to high variance and poor alignment [2]. Gradient mismatch leads to unstable training and collapse.

(5) Increased memory and runtime. Distillation requires caching teacher predictions and computing ODE steps between time pairs. This is particularly prohibitive for large 3D models and long training horizons, where memory and runtime grow linearly with Δt supervision steps [8, 17, 45].

(6) Empirical failure in 3D. Previous 3D works report degraded geometric quality and instability when using teacher–student CMs [5, 8]. Our own ablation studies confirm this: teacher-based losses fail to capture fine-grained geometry, leading to blurred, collapsed, or misaligned reconstructions. Quantitatively, they underperform on both Chamfer and Earth Mover's metrics (see Section 5).

Conclusion. Due to the absence of semantic alignment, point permutation invariance, temporal instability, and computational overhead, teacher–student consistency is *incompatible* with point cloud generation. This motivates our **teacher-free design**, which replaces teacher-based supervision with: (a) a flow-matching regression to the known analytic TrigFlow direction (Eq. 11), and (b) a permutation-invariant Chamfer loss (Eq. 13), yielding a stable, efficient, and geometry-aware training objective.

Table 8. Noise schedule ablation (ShapeNet [4] *chair*, $S = 2$).

Schedule	1-NNA (CD)↓	MMD (EMD)↓
Linear FM	58.2	2.73
Cosine-DDPM	57.9	2.65
TrigFlow (ours)	**54.3**	**2.42**

10. Additional Experimental Results

10.1. Ablation Results

We report extended ablation experiments to complement the discussion in Sec. 5.3 of the main paper. These results further disentangle the contributions of our loss formulation, noise schedule, and sampling strategy.

Loss components. Table 6 compares FM, Chamfer reconstruction, and their combination. FM only training produces diverse but geometrically imprecise outputs, as indicated by higher EMD and Jensen–Shannon Divergence [1] (JSD) values. Chamfer-only training improves fidelity but collapses to limited modes, yielding poor coverage [1] (COV). The joint FM+Chamfer objective strikes the best balance, achieving the lowest 1-NNA and Minimum Matching Distance [1] (MMD) along with the highest COV, confirming the complementarity of reconstruction and analytic supervision.

Noise schedule. Table 8 evaluates different forward noise schedules on ShapeNet [4] class: *chair*. Both linear and cosine schedules lead to degraded 1-NNA and MMD compared to our proposed TrigFlow. TrigFlow provides more stable training and improved geometric fidelity in the few-step regime, validating its use as the default schedule.

Sampling steps. Table 7 reports the effect of varying the number of inference steps S on ShapeNet [4] class: *car*.

Performance improves markedly when increasing from $S = 1$ to $S = 2$, with diminishing returns beyond. Although $S = 4$ achieves only marginal gains, it doubles the runtime. Interestingly, Heun's solver provides slightly better single-step results than Euler at minimal extra cost. In general, $S = 2$ with Euler offers the best trade-off between accuracy and efficiency.

Summary. Overall, the ablation results confirm that Chamfer reconstruction complements analytic FM by improving fidelity without reducing diversity, that TrigFlow is essential for stable one- and two-step sampling, and that most performance gains are already achieved with $S = 2$ steps. These findings support the design of ConTiCoM-3D and demonstrate that its contributions are robust and largely orthogonal to backbone size or external teacher supervision.

10.2. Further Qualitative Results

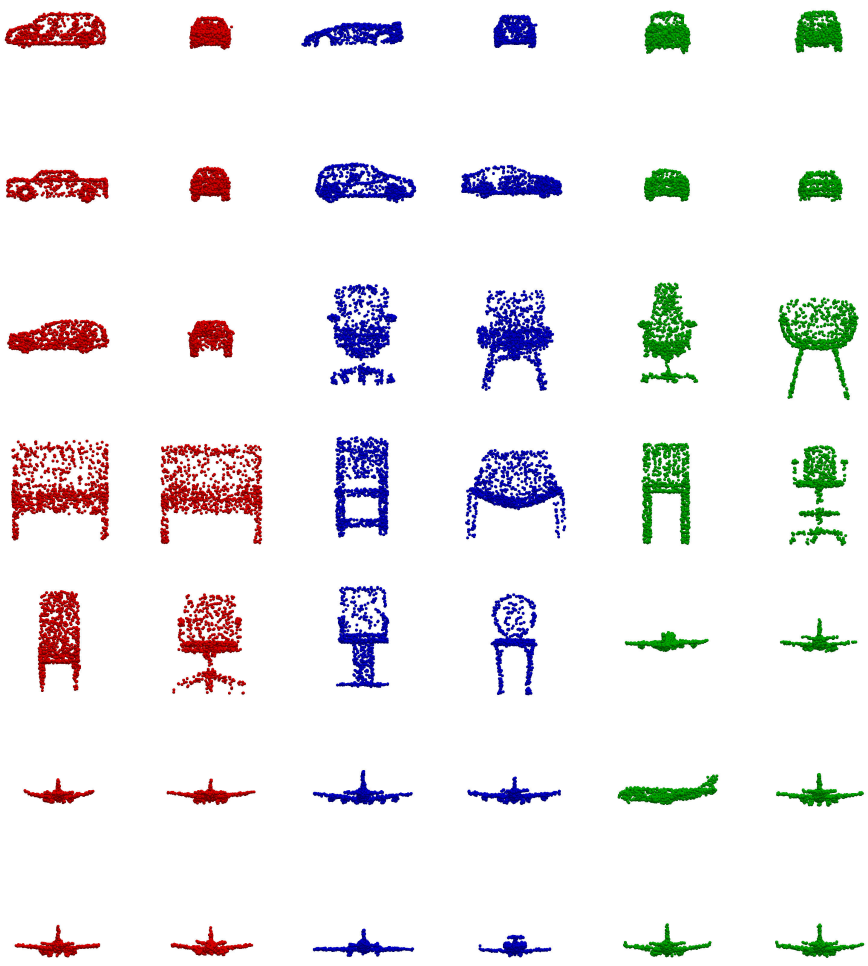

Figure 4. Further qualitative results with ConTiCoM-3D (S=1) on ShapeNet dataset