VAP-RTS Improves Training and Closed-Loop Feature Generation

Anonymous Supplementary Material for Rebuttal

1 Summary

We test **VAP-RTS** (VAP-guided Reducible-error Time Sampling), a time-sampling intervention derived from the paper's distinction between irreducible velocity ambiguity and reducible excess risk. In a fine-conditioned flow-matching task over DINOv2 features extracted from CIFAR-100 images, VAP-RTS improves velocity-prediction accuracy. Compared with the baseline using a uniform sampler, VAP-RTS reduces the time-averaged velocity squared-error loss by **6.49%**; at the same training step, the reduction is **6.78%**. All five paired seeds favor VAP-RTS in both comparisons. When the resulting vector fields are deployed as closed-loop ODE generators, VAP-RTS reduces global DINO feature FD by **23.6%** and improves precision, density, coverage, and fine-label accuracy in all five seeds. The supported claim is improvement of the uniform-time training functional and multiple closed-loop feature-generation metrics—not of every discrepancy or diversity metric.

2 From VAP theory to VAP-RTS

The paper defines the Velocity Ambiguity Profile $\mathcal{A}_v(t)$ as the time-local Bayes risk of velocity prediction and proves the exact decomposition

$$L_N(t) = \underbrace{\mathcal{A}_v(t)}_{\text{irreducible ambiguity}} + \underbrace{R(t;N)}_{\text{reducible excess risk}} .$$

This decomposition gives VAP-RTS its central principle: training time should be allocated according to *reducible risk*, not raw local loss. A large loss may simply reflect irreducible velocity ambiguity, in which case sampling that time more often need not help. VAP-RTS instead seeks the times at which the current learner remains improvable and concentrates training there.

The training flow begins with a high-capacity *reference model*, whose time-local loss provides a stable empirical reference $L_{\text{ref}}(t) = \mathcal{A}_v(t) + R_{\text{ref}}(t)$. Because the true ambiguity profile is unobserved, this model is not treated as the Bayes floor or as a teacher. Independently, a *source model* with the downstream target architecture is trained only through an early phase under uniform time sampling. Its loss profile records where a target-capacity learner initially falls furthest above the reference. Evaluating the two roles under the same held-out conditions yields

$$\widehat{R}_{\text{rel}}(t) = [L_{\text{source}}(t) - L_{\text{ref}}(t)]_+ = [R_{\text{source}}(t) - R_{\text{ref}}(t)]_+ .$$

This positive gap is a relative reducible-risk proxy: it emphasizes where the early source learner underperforms the empirical reference, rather than where both models incur unavoidable ambiguity. The gap is then converted once into a frozen VAP-RTS time sampler according to $q(t) \propto \widehat{R}_{\text{rel}}(t)^4$. Finally, fresh *target models* are trained from scratch: the guided targets use that sampler, while their paired baselines retain uniform time sampling. No model parameters or optimizer state pass between roles; the frozen sampler is the only source/reference-to-target link. The reference defines the comparison profile, the source identifies where learning lags, and the targets test whether reallocating training time improves the learned vector field. This is a theory-motivated empirical procedure, not a theorem-optimal universal scheduler.

3 Experimental design and role isolation

We extract frozen DINOv2 ViT-S/14 final-LayerNorm CLS features ($d = 384$) from the official CIFAR-100 split. A label-free transform, fitted only on the 50,000 training features, subtracts the train mean and divides by one global RMS scalar. For calibrated feature z, Gaussian endpoint ϵ, fine label c, and time t,

$$x_t = (1-t)z + t\epsilon, \qquad u_t = \epsilon - z, \qquad \mathcal{L} = \mathbb{E}\|v_\theta(x_t, t, c) - u_t\|_2^2.$$

Table 1: *Compact model and training hyperparameters. Both are five-affine-layer GELU MLPs; source proxies and paired targets share target capacity but are independently trained.*

Setting	High-capacity full-data reference	Source proxies / paired targets
Network	$432 \to 2048 \times 4 \to 384$ (14.26M parameters)	$432 \to 512 \times 4 \to 384$ (1.21M parameters)
Optimizer	Adam, constant 10^{-3}, batch 512	Adam, constant 3×10^{-4}, batch 512
Training t	continuous $\mathcal{U}(0.01, 0.99)$	source: uniform; targets: uniform / VAP-RTS

Reference seeds 500–504, source-proxy seeds 520–529, and paired target seeds 540–544 are mutually disjoint. The four evaluation banks are also pairwise disjoint within the experiment; each contains 1,000 test anchors, exactly 10 per fine class:

Table 2: *Pre-specified role-separated anchor banks. Positions are within each fine class after the deterministic classwise shuffle. All six pairwise intersections are empty.*

Bank	Per-class positions	Exclusive role
Reference selection	20–29	select reference checkpoints
Source proxy	10–19	evaluate source and selected references
Target selection	30–39	select the seed-mean common step
Target confirmation	0–9	evaluate after selection; no reselection

Each reference is selected only on the reference-selection bank and then re-evaluated on the source-proxy bank without reselection. The proxy uses that same source bank for the matched source–reference difference, but neither source nor reference bank overlaps either target bank. The anchor indices were fixed before model training and were not selected, removed, or reordered using per-anchor losses or class-specific outcomes. The confirmation bank is therefore isolated from training, proxy construction, and checkpoint selection.

The resulting VAP-RTS distribution has positive mass in every one of 64 seamless bins, ESS 47.17, and density ratio $q/u \in [0.223, 2.215]$ (Fig. 1). The paired target runs share the exact training subset/order, initialization, batch stream, Gaussian endpoint-noise stream, and evaluation noise; only the training-time t sampler changes.

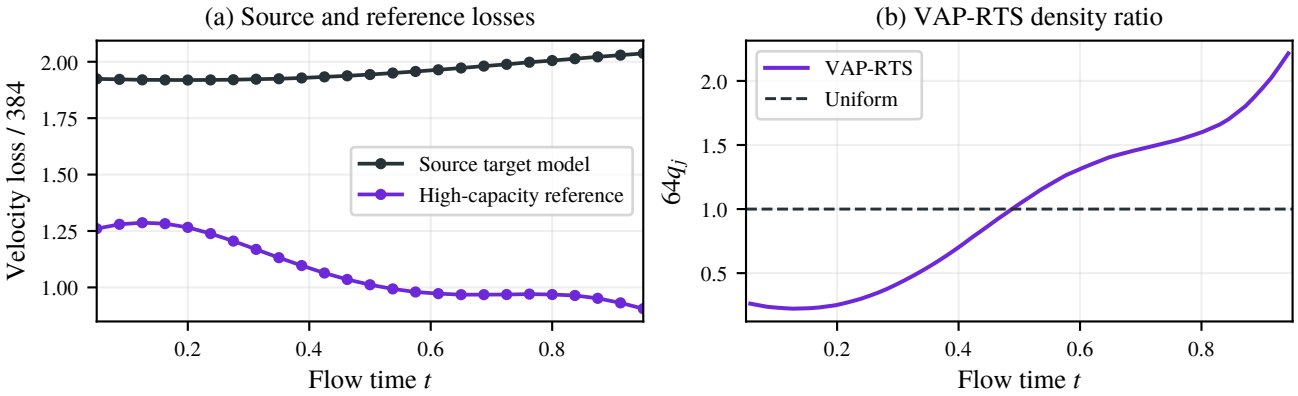

Figure 1: *VAP-RTS construction. Left: independently trained source and high-capacity reference profiles. Right: the resulting VAP-RTS density relative to uniform sampling.*

4 Uniform-time target-selection result

Figure 2 reports trajectories from the target-selection bank. Its test anchors are unseen in training and disjoint from the source and reference banks. The best seed-mean checkpoints are step 8,500 for uniform and step 11,500 for VAP-RTS. VAP-RTS reduces loss by **6.49%** at the methods' respective best checkpoints and by **6.78%** in the strict same-step control at step 11,500, with **5/5** paired seeds improving in both comparisons. Evaluation is uniform in time, so the gain is not produced by evaluating under the intervention density.

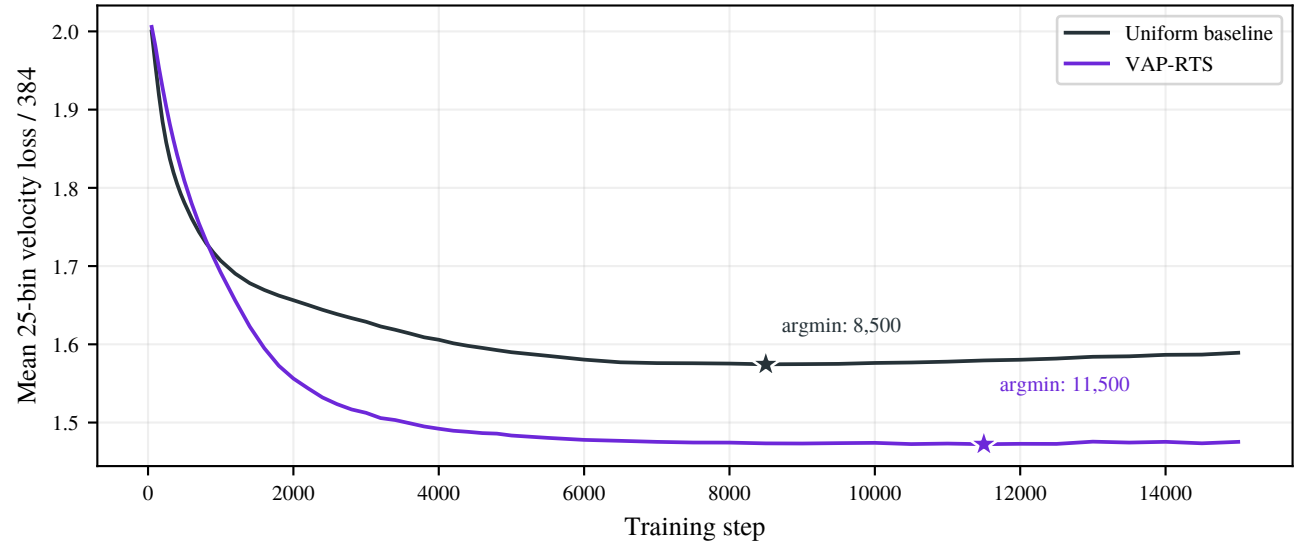

Figure 2: *Target-selection-bank 25-bin uniform-time trajectories (mean ± SEM over five paired seeds). Stars mark seed-mean common-step argmins.*

5 Closed-loop feature generation

We deploy one common checkpoint per method (uniform: step 8,500; VAP-RTS: step 10,500) and integrate $dx/dt = v_\theta(x, t, c)$ backward from $x(1) \sim \mathcal{N}(0, I_{384})$ to $t = 0$ using 100-step fixed-step Heun. For each seed, class, and within-class index, the two methods use byte-identical initial noise. Each of the five models generates 60 features for each of 100 fine classes (30,000 generated features per method); metrics use a balanced 6,000-feature held-out test bank, disjoint from all four role banks in Table 2.

Table 3: *Closed-loop endpoint metrics in the full 384-D calibrated feature space. Effects are VAP-RTS minus uniform; intervals are paired 95% t_4 CIs.*

Metric	Uniform	VAP-RTS	Paired effect	95% CI
Global feature FD ↓	33.333	**25.476**	**-7.858**	[-8.683,-7.033]
Precision $k = 3$ ↑	0.151	**0.690**	**+0.539**	[+0.527,+0.551]
Density $k = 5$ ↑	0.122	**1.226**	**+1.104**	[+1.052,+1.156]
Coverage $k = 5$ ↑	0.234	**0.728**	**+0.494**	[+0.476,+0.511]
Fine accuracy ↑	0.791	**0.859**	**+0.0676**	[+0.055,+0.080]
Fine cross-entropy ↓	1.094	**0.626**	**-0.468**	[-0.520,-0.417]

Every effect in Table 3 has the favorable direction in **5/5 paired seeds**. The global FD effect is unchanged to the reported precision for 25, 50, 100, and 200 Heun steps, arguing against a coarse-integration explanation. Generated features also have smaller nearest-neighbor distance to target-train and held-out-test features, higher nearest-train label agreement, and zero exact float32 duplicates.

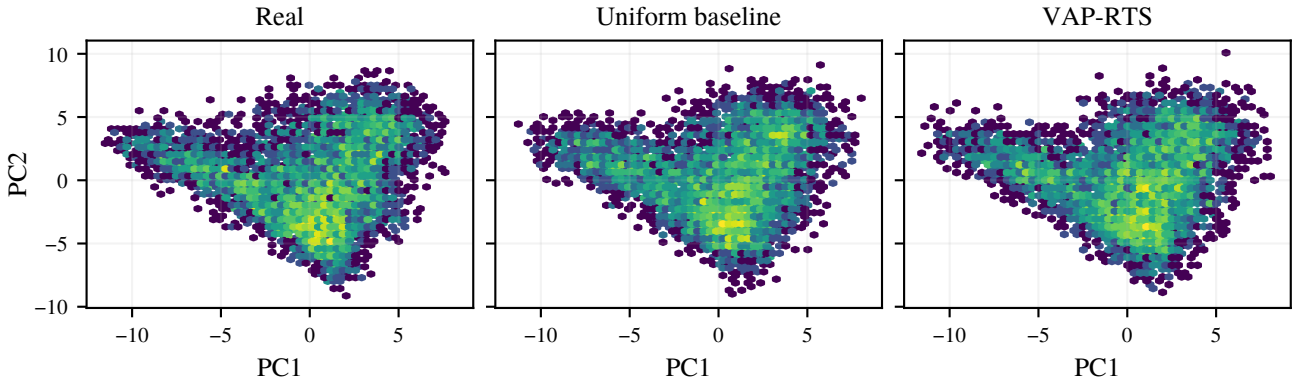

Figure 3: *Real and generated endpoints under one PCA fitted once on official calibrated training features. This is qualitative support; all reported metrics are computed in 384 dimensions.*

6 Claim boundary

The result exhibits a fidelity–diversity trade-off. Macro conditional KID is 0.1988 for uniform and 0.2062 for VAP-RTS (paired interval crosses zero), while improved recall decreases from 0.7221 to 0.2214 and effective covariance rank from 47.25 to 40.14. We therefore claim improvement on the uniform-time selection objective and on **multiple** closed-loop feature metrics, not on every distribution metric. The VAP-RTS family/checkpoint used for deployment is post-selected from the target-selection trajectories; these deployment results are conditional on that fixed selection procedure rather than an independent confirmation experiment. The study is feature-space only: feature FD is not image FID, no pixel realism claim is made, and the statistical unit is the paired training seed rather than the 30,000 generated samples.