LEARNING A FLOW TO SELF-SUPERVISED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Explicit geometric references offer a direct way to structure self-supervised representations. Existing adversarial distribution-matching formulations, however, require costly encoder–critic optimization (Jiao et al., 2026). We introduce Flow-Based Distribution Matching (FBDM), a non-adversarial framework that shapes the encoder toward an explicit geometric reference via spherical conditional velocity regression during pretraining, then discards the learned velocity field for downstream evaluation. An ETF-inspired reference allows its number of components K' to exceed the auxiliary flow dimension $d^\star$ while retaining structured geometric separation. We assign both augmented views of each image to the same target, while limiting how many images each reference center can receive. An explicit alignment loss further pulls the two views' representations closer together. Experiments across benchmarks ranging from CIFAR to ImageNet show that FBDM achieves performance nearly on par with DM and remains competitive with existing SSL methods. Matched training-cost comparisons show a 1.48–1.83$\times$ speedup over DM with a negligible increase in GPU memory usage. We also provide a theoretical explanation for the usefulness of the learned representations: under stated conditions, we bound the downstream misclassification rate in terms of the FBDM pretraining loss.

1 INTRODUCTION

Self-supervised learning (SSL) learns transferable representations without labels, but must prevent collapse: perfect view agreement is useless if all images map to the same vector. Contrastive, predictor-based, and redundancy-reduction methods address this through negative pairs, asymmetry, or statistical constraints (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021; Hua et al., 2021; Ermolov et al., 2021; Bardes et al., 2022). Recent work continues to develop these objectives and their analysis (Huang et al., 2024; Duan et al., 2025; Siméoni et al., 2025; Zhang et al., 2026; Luthra et al., 2026). These approaches largely specify what representations should *not* become, leaving their desired global distribution implicit.

Distribution Matching (DM) offers a generative perspective: it prescribes an explicit reference distribution ν and trains the encoder f by minimizing $W_1(f_\# P_\mathcal{A}, \nu)$, where W_1 denotes the Wasserstein-1 distance and $f_\# P_\mathcal{A}$ is the encoded augmentation distribution. In practice, this distance is estimated by an adversarial critic (Jiao et al., 2026). Two restrictions remain. Its coordinate reference requires $K' \leq d^\star$, coupling the number of latent templates to the matching dimension. Its critic-based objective requires opposing updates: the critic increases the estimated discrepancy while the encoder decreases it (Arjovsky et al., 2017). Tracking the evolving encoder requires multiple critic steps per encoder update and careful hyperparameter tuning. Can we retain explicit distribution design without the adversarial critic, while allowing more templates in a compact flow space?

Our answer is FBDM. Generative learning is itself a distribution-transport problem, and flow matching provides analytic velocity targets instead of critic-estimated distances (Lipman et al., 2023; Liu et al., 2023). FBDM uses this tool for representation learning rather than sample generation: its flow objective guides the encoder toward an explicit geometric reference. The auxiliary velocity field provides a training signal but is discarded downstream, without repeated critic optimization. An ETF-inspired simplex-spectral reference permits $K' > d^\star$; a two-view consensus assigns each image a shared destination, and an explicit starting-point penalty reinforces alignment. Figure 1

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.

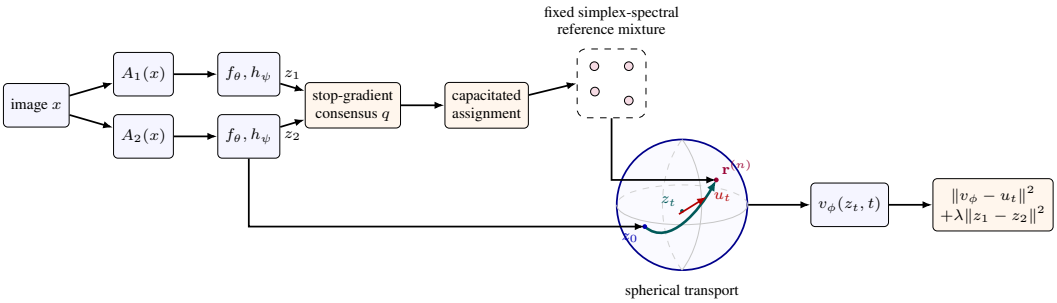

Figure 1: **FBDM in one view.** Two augmented views are encoded on the unit sphere. Their stop-gradient consensus selects a capacity-constrained reference component and a locally perturbed target $\mathbf{r}^{(n)}$. Conditional flow matching regresses the target velocity u_t along the resulting geodesic z_t, while an explicit start-point loss aligns the views. Only the backbone is retained downstream.

summarizes the pipeline. Training needs neither critic maximization nor a numerical ODE solve; only the backbone is retained downstream. Experiments across benchmarks ranging from CIFAR to ImageNet show that this approach learns competitive representations, while matched training-cost comparisons demonstrate a 1.48–1.83× speedup over DM with modest additional GPU memory (Section 4).

Our population theory links the full FBDM loss to classification from the initial representation. Under augmentation-quality, regularity, and transport-identifiability assumptions, Theorem 1 bounds nearest-centroid error by an augmentation-dependent residual plus a square-root loss term, in a positive-margin, small-loss regime. Velocity regression controls terminal Wasserstein error; identifiability and ODE stability transfer class separation to the initial representation, while alignment and augmentation quality control within-class variation.

Our contributions are:

- An ETF-inspired reference permits $K' > d^\star$ with structured geometric separation, decoupling reference granularity from the auxiliary flow dimension.
- Building on this reference, FBDM replaces critic-based distribution matching with velocity regression. It achieves comparable performance to DM and competitive SSL results across benchmarks ranging from CIFAR to ImageNet, with 1.48–1.83× higher training throughput and about 0.50 GiB additional peak allocated memory under matched hardware.
- A conditional population guarantee bounds downstream misclassification by the full FBDM loss and augmentation quality, for a nearest-centroid classifier on the initial normalized representation.

2 Related Work

SSL theory studies latent-class transfer, alignment and uniformity, spectral structure, and downstream generalization (Saunshi et al., 2019; Wang & Isola, 2020; HaoChen et al., 2021; 2022; Huang et al., 2023; Duan et al., 2025). DM connects representation quality to explicit Wasserstein matching (Jiao et al., 2026). Flow matching instead learns distributional transport by velocity regression (Lipman et al., 2023; Fukumizu et al., 2025), including geodesic constructions on manifolds (Chen & Lipman, 2024). REPA reports that generative diffusion features lag dedicated SSL features on classification probes, and uses pretrained features to improve generation (Yu et al., 2025). Self-Flow learns representations within multimodal synthesis (Chefer et al., 2026). SenFlow (Ukita & Okita, 2026) jointly trains an encoder and a conditional flow generator in an autoencoder-like design, targeting both recognition and generation. Generation remains an objective in all three.

FBDM instead uses reference-directed velocity regression to train an encoder for downstream recognition, without generating data or using a pretrained teacher. Appendix F discusses additional SSL methods, probability-flow ODE theory, and frame geometry.

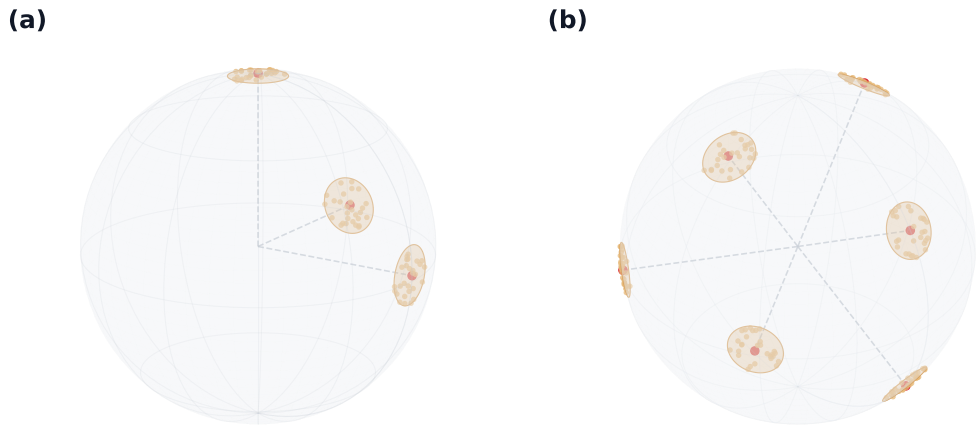

Figure 2: **From a coordinate reference to an overcomplete spherical reference.** (a) Jiao et al. (2026) assigns at most one template to each orthogonal coordinate direction, which requires $K' \leq d^\star$. (b) Our construction permits $K' > d^\star$ and places more reference components on the unit sphere in the $d^\star$-dimensional flow space. The pale point clouds depict local samples $(\mathbf{c}_k + \varepsilon\boldsymbol{\eta})/\|\mathbf{c}_k + \varepsilon\boldsymbol{\eta}\|_2$ around each center, with $\boldsymbol{\eta} \sim \mathrm{Unif}(\mathbb{S}^{d^\star-1})$; they are not additional centers. The illustration is schematic, while Proposition 1 gives the precise ETF/Welch geometry.

3 Flow-Based Distribution Matching

Notation. Vectors and matrices are bold; $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{R}$ denote random vectors. We reserve d for the input-image dimension and $d^\star$ for the auxiliary flow-space dimension. Write $[N] = \{1, \ldots, N\}$, $\mathbb{S}^{q-1} = \{\mathbf{u} \in \mathbb{R}^q : \|\mathbf{u}\|_2 = 1\}$, and $\mathbf{I}_q, \mathbf{1}$ for the identity matrix and all-ones vector. For a set $\mathcal{S}$, $\mathrm{Unif}(\mathcal{S})$ is its uniform law; sg is stop-gradient and $\hookrightarrow$ an injective map.

3.1 Representations and a Flexible Reference

Given unlabeled images $\mathcal{D}_N = \{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$, each iteration draws B images. For each $\mathbf{x}^{(n)}$, draw two views independently from the augmentation kernel $\mathcal{A}(\cdot \mid \mathbf{x}^{(n)})$. Since assignment compares directions, normalize the backbone–projector output:

$$\mathbf{z}_j^{(n)} = \frac{h_\psi(f_\theta(\mathbf{x}_j^{(n)}))}{\|h_\psi(f_\theta(\mathbf{x}_j^{(n)}))\|_2} \in \mathbb{S}^{d^\star-1}, \qquad j \in \{1, 2\}. \tag{1}$$

We now specify a destination law, pair views with its components, and construct the resulting transport signal.

More components without a larger flow space. DM uses $\mathbf{c}_k^{\mathrm{DM}} = s_k \mathbf{e}_{\pi(k)}$, where $\mathbf{e}_i$ is the ith standard basis vector in $\mathbb{R}^{d^\star}$, with independent $s_k \sim \mathrm{Unif}(\{-1, 1\})$ and $\pi : [K'] \hookrightarrow [d^\star]$ (Jiao et al., 2026). Thus $K' \leq d^\star$: discovering more latent groups requires increasing the matching dimension $d^\star$. To allow finer template granularity without increasing this dimension, we instead construct K' unit centers $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{K'}$ with $K' > d^\star$. Figure 2 illustrates this decoupling.

From an ideal geometry to a compact reference. To give these components a geometrically separated arrangement, we use an equiangular tight frame (ETF) as the ideal template: it minimizes the largest absolute inner product between distinct unit centers (Appendix A.1). However, an exact ETF need not exist for an arbitrary pair $(K', d^\star)$. We therefore use a Gram-first construction: for the guaranteed pair $(K', K' - 1)$, we prescribe the ideal simplex Gram matrix

$$\mathbf{G}_\Delta = \frac{K'}{K' - 1}\left(\mathbf{I}_{K'} - \frac{\mathbf{1}\mathbf{1}^\top}{K'}\right) \in \mathbb{R}^{K' \times K'}. \tag{2}$$

Its diagonal entries are 1 and off-diagonal entries are $-1/(K'-1)$, directly encoding unit norms and equiangularity. Appendix A.2 verifies that $\mathbf{G}_\Delta$ is positive semidefinite with rank $K'-1$ and spectrally realizes an exact ETF of K' vectors in $\mathbb{R}^{K'-1}$. We then project this ideal geometry to the target matching dimension $2 \le d^\star < K'-1$.

Let $\mathbf{U}_{d^\star}$ and $\mathbf{\Lambda}_{d^\star}$ collect the leading $d^\star$ orthonormal eigenvectors and eigenvalues. The projected coordinate matrix is

$$\mathbf{Y} = \mathbf{U}_{d^\star} \mathbf{\Lambda}_{d^\star}^{1/2} \in \mathbb{R}^{K' \times d^\star}. \tag{3}$$

Lemma 2 shows that (3) minimizes the total squared error in the simplex's pairwise inner products before normalization. We then obtain spherical centers by normalizing each row:

$$\mathbf{c}_k = \frac{\mathbf{Y}_{k,:}}{\|\mathbf{Y}_{k,:}\|_2}, \qquad k \in [K'].$$

The retained eigenspace is chosen to give nonzero rows and distinct normalized directions. Since normalization only rescales each nonzero row by a positive scalar, it preserves the directions and pairwise angles of the projected rows. The construction recovers the exact simplex at $d^\star = K'-1$ and an ETF-inspired reference in lower dimensions. Appendix A.2 gives the eigenpair definitions and Algorithm 2 summarizes the construction.

To retain local variation rather than force all images in a component to the same point, perturb its center by a spherical direction, with $0 \le \varepsilon < 1$:

$$Q_k^\varepsilon = \mathrm{Law}\left(\frac{\mathbf{c}_k + \varepsilon \boldsymbol{\eta}}{\|\mathbf{c}_k + \varepsilon \boldsymbol{\eta}\|_2} \right), \qquad \nu_\varepsilon = \frac{1}{K'} \sum_{k=1}^{K'} Q_k^\varepsilon, \quad \boldsymbol{\eta} \sim \mathrm{Unif}(\mathbb{S}^{d^\star - 1}). \tag{4}$$

Here K' controls template granularity and ε local spread; components are label-free templates, not prescribed ground-truth classes.

3.2 Two-View Capacitated Assignment

Flow matching must pair each view with a destination. To align a positive pair implicitly, we choose one shared destination using its detached consensus:

$$\mathbf{q}^{(n)} = \frac{\mathrm{sg}(\mathbf{z}_1^{(n)}) + \mathrm{sg}(\mathbf{z}_2^{(n)})}{\|\,\mathrm{sg}(\mathbf{z}_1^{(n)}) + \mathrm{sg}(\mathbf{z}_2^{(n)})\|_2}. \tag{5}$$

Detachment keeps the combinatorial assignment outside the gradient path. The normalized sum is defined for non-antipodal views.

A directly balanced construction assumes $K' \mid B$, meaning that K' divides B, and allocates $m_0 = B/K'$ images to each center. However, minibatches need not be semantically balanced, and fixed per-center quotas may force images sharing similar semantics toward incompatible components. To relax this fixed quota, we use an epoch-indexed capacity $m_{\mathcal{D}_N}(e) > B/K'$, where e denotes the training epoch, and select only B of the available slots:

$$\mathcal{P}_{\mathcal{D}_N}(e) = [K'] \times [m_{\mathcal{D}_N}(e)], \qquad \kappa : [K'] \times \mathbb{N} \to [K'], \quad \kappa((k,\ell)) = k. \tag{6}$$

The injective assignment

$$\alpha^* \in \arg \max_{\alpha : [B] \hookrightarrow \mathcal{P}_{\mathcal{D}_N}(e)} \sum_{n=1}^{B} (\mathbf{q}^{(n)})^\top \mathbf{c}_{\kappa(\alpha(n))} \tag{7}$$

caps occupancy without enforcing equality. Draw $\boldsymbol{\eta}^{(n)} \sim \mathrm{Unif}(\mathbb{S}^{d^\star - 1})$ and set

$$\mathbf{r}^{(n)} = \frac{\mathbf{c}_{\kappa(\alpha^*(n))} + \varepsilon \boldsymbol{\eta}^{(n)}}{\|\mathbf{c}_{\kappa(\alpha^*(n))} + \varepsilon \boldsymbol{\eta}^{(n)}\|_2}. \tag{8}$$

The resulting $\mathbf{r}^{(n)}$ is the shared endpoint toward which both $\mathbf{z}_j^{(n)}$ are transported along the velocity field; sharing it across the two views implements implicit alignment. The capacity schedule controls per-center availability; Appendix D.3, Table 5, gives the dataset-specific choices.

3.3 Spherical Transport and Its Learning Signal

Because representations and destinations lie on the sphere, a shortest great-circle path is a natural interpolation (Chen & Lipman, 2024). For non-antipodal $\mathbf{z}, \mathbf{r}$, let $\omega = \arccos(\mathbf{z}^\top \mathbf{r})$ and define progress $s \in [0, 1]$ by

$$\boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(s) = \frac{\sin((1-s)\omega)}{\sin \omega}\mathbf{z} + \frac{\sin(s\omega)}{\sin \omega}\mathbf{r}. \tag{9}$$

The $\omega = 0$ case is understood by continuity. Appendix B.1 derives the spherical path and its tangent velocity. A comparison with linear interpolation is provided in Appendix E.2.

Fix a training epoch e and suppress its dependence in the notation. Assume that the assignment in Section 3.2 has produced the representation–endpoint pairs, and denote their induced coupling by $\Pi_f = \mathrm{Law}(\mathbf{Z}_0, \mathbf{R})$. For $(\mathbf{Z}_0, \mathbf{R}) \sim \Pi_f$, a continuously differentiable increasing clock $s(t)$ with $s(0) = 0, s(1) = 1$ defines

$$\mathbf{Z}_t = \boldsymbol{\gamma}_{\mathbf{Z}_0,\mathbf{R}}(s(t)), \qquad p_t = \mathrm{Law}(\mathbf{Z}_t). \tag{10}$$

Its analytic velocity is the signal to be regressed:

$$\mathbf{U}_t = \frac{\mathrm{d}\mathbf{Z}_t}{\mathrm{d}t} = \dot{s}(t)\, \partial_s \boldsymbol{\gamma}_{\mathbf{Z}_0,\mathbf{R}}(s)|_{s=s(t)}\,. \tag{11}$$

The conditional mean $\boldsymbol{v}^\star(\mathbf{z}, t) = \mathbb{E}[\mathbf{U}_t \mid \mathbf{Z}_t = \mathbf{z}]$ realizes the distributional path through

$$\partial_t p_t(\mathbf{z}) + \nabla \cdot (p_t(\mathbf{z})\boldsymbol{v}^\star(\mathbf{z}, t)) = 0. \tag{12}$$

Here $\nabla\cdot$ is divergence; when p_t has no ambient density, (12) is understood in the standard weak sense. Conditional expectation connects path construction to velocity regression (Lipman et al., 2023):

$$\mathcal{R}_{\mathrm{CFM}}(\phi; q_T) = \frac{1}{d^\star}\mathbb{E}_{\Pi_f,T}\left[\|\boldsymbol{v}_\phi(\mathbf{Z}_T, T) - \mathbf{U}_T\|_2^2\right], \qquad T \sim q_T, \tag{13}$$

where T is independent of the endpoints.

Time reparameterization and uniform-progress sampling. For semantic encoding, reaching an appropriate reference component matters more than precise positioning within it. We therefore choose clocks with decreasing $\dot{s}(t)$. Through (11), this strengthens component-directed motion early and attenuates within-component refinement late without changing the path. For example, the rational schedule used on several datasets has $\dot{s}(t) = (1 + a)/(1 + at)^2$, which decreases for $a > 0$. The reported rational and exponential choices appear in Table 6 and follow the distance-scheduling perspective of Chen & Lipman (2024). Uniformly sampling t would overrepresent the slow late segment, so inverse-transform sampling instead draws $S_j^{(n)} \sim \mathrm{Unif}([0, 1])$ and sets

$$T_j^{(n)} = s^{-1}(S_j^{(n)}). \tag{14}$$

This gives $q_T(t) = \dot{s}(t)$ and uniform progress coverage: early states carry stronger targets rather than being sampled more often. The network takes $(\mathbf{Z}_j^{(n)}, T_j^{(n)})$ and predicts $\mathbf{U}_j^{(n)}$, where

$$\mathbf{Z}_j^{(n)} = \boldsymbol{\gamma}_{\mathbf{z}_j^{(n)},\mathbf{r}^{(n)}}(S_j^{(n)}), \qquad \mathbf{U}_j^{(n)} = \dot{s}(T_j^{(n)})\partial_s \boldsymbol{\gamma}_{\mathbf{z}_j^{(n)},\mathbf{r}^{(n)}}(S_j^{(n)}). \tag{15}$$

3.4 Objective, Updates, and Inference

A shared endpoint provides only implicit alignment. The no-alignment diagnostic in Appendix E.3 motivates an explicit penalty on the starting representations:

$$\widehat{\mathcal{R}}_{\mathrm{FM},B} = \frac{1}{2Bd^\star}\sum_{n=1}^{B}\sum_{j=1}^{2}\|\boldsymbol{v}_\phi(\mathbf{Z}_j^{(n)}, T_j^{(n)}) - \mathbf{U}_j^{(n)}\|_2^2, \tag{16}$$

$$\widehat{\mathcal{R}}_{\mathrm{align},B} = \frac{1}{Bd^\star}\sum_{n=1}^{B}\|\mathbf{z}_1^{(n)} - \mathbf{z}_2^{(n)}\|_2^2, \tag{17}$$

$$\widehat{\mathcal{R}}_B = \widehat{\mathcal{R}}_{\mathrm{FM},B} + \lambda\widehat{\mathcal{R}}_{\mathrm{align},B}, \qquad \lambda \geq 0. \tag{18}$$

Algorithm 1 Core FBDM pretraining loop.

Require: $\mathcal{D}_N$, $\mathcal{A}$, $\mathcal{C}$, ε, λ, $m_{\mathcal{D}_N}(e)$, $s(t)$; networks f_θ, h_ψ, $\boldsymbol{v}_\phi$

1: **for** each epoch e and minibatch $\mathcal{B} \subset \mathcal{D}_N$ **do**
2: Draw two views and compute $\mathbf{z}_j^{(n)}$ by (1)
3: Form $\mathbf{q}^{(n)}$ and solve the capacitated assignment (7)
4: Draw one endpoint $\mathbf{r}^{(n)}$ by (8) and share it across both views
5: Draw $S_j^{(n)}$, obtain $T_j^{(n)}$, and form $(\mathbf{Z}_j^{(n)}, \mathbf{U}_j^{(n)})$ by (14)–(15)
6: Evaluate (18) and update (θ, ψ, ϕ)
7: **return** f_θ

The empirical risk averages this objective over random minibatches and augmentation, perturbation, and progress randomness: $\widehat{\mathcal{R}}_N = \mathbb{E}_{\mathcal{I},\xi}[\widehat{\mathcal{R}}_B]$. Here $\mathcal{I}$ is a uniform size-B subset of $[N]$ and ξ collects the other sampling variables. Since assignment couples the samples, this is a batch-level expectation; (111) gives the explicit sampling law.

Algorithm 1 uses detached assignments. The reported ResNet-18 runs attenuate the FM gradient entering the encoder and projector, allowing the velocity field to track their evolving geometry. Appendix D.2 specifies these backward-only updates; they are not ordinary gradients of the unmodified scalar risk. After pretraining, only the 512-dimensional backbone feature $f_\theta(\mathbf{x})$ is used for linear and 5-NN evaluation.

4 Experiments

4.1 Accuracy and Representation Quality

We evaluate CIFAR-10, CIFAR-100, STL-10, and Tiny ImageNet with frozen ResNet-18 features using linear classification and cosine 5-NN (Table 1), and report ImageNet-1K separately under ResNet-50 pretraining for 100 epochs at global batch size 512 (Table 2). Full configurations are given in Appendix D.3.

Table 1: Top-1 accuracy (%) on CIFAR-10, CIFAR-100, STL-10, and Tiny ImageNet with ResNet-18. *Linear* evaluates the frozen representation; 5-NN uses $k = 5$. Non-FBDM results except DM follow the published benchmark (Weng et al., 2022); DM results are taken from its source paper (Jiao et al., 2026).

	CIFAR-10		CIFAR-100		STL-10		Tiny ImageNet	
Method	Linear	5-NN	Linear	5-NN	Linear	5-NN	Linear	5-NN
FBDM	92.37	89.68	66.59	56.74	89.79	86.23	48.55	33.01
DM (Jiao et al., 2026)	92.11	89.17	67.71	56.18	90.22	85.51	–	
Barlow Twins (Zbontar et al., 2021)	88.51	86.53	65.78	55.76	88.36	83.71	47.44	32.65
SimCLR (Chen et al., 2020)	91.80	88.42	66.83	56.56	90.51	85.68	48.84	32.86
BYOL (Grill et al., 2020)	91.73	89.45	66.60	56.82	91.99	88.64	51.00	36.24
SimSiam (Chen & He, 2021)	90.51	86.82	66.04	55.79	88.91	84.84	48.29	34.21
Shuffled-DBN (Hua et al., 2021)	90.45	88.15	66.07	56.97	89.20	84.51	48.60	32.14
Zero-ICL (Zhang et al., 2022)	88.12	86.64	61.91	53.47	86.35	82.51	46.25	32.74
VICReg (Bardes et al., 2022)	90.32	88.41	66.45	56.78	90.78	85.72	48.71	33.35
W-MSE (Ermolov et al., 2021)	91.55	89.69	66.10	56.69	90.36	87.10	48.20	34.16
CW-RGP (Weng et al., 2022)	91.92	89.54	67.51	57.35	90.76	87.34	49.23	34.04

Table 2: ImageNet-1K linear-evaluation top-1 accuracy (%) with ResNet-50 after 100 epochs of pretraining at global batch size 512. The SimCLR value is the reproduction reported by `AndrewAtanov/simclr-pytorch`.

Method	Linear
FBDM	61.41
SimCLR (Chen et al., 2020)	60.14

Across the benchmark suite from CIFAR to ImageNet, FBDM achieves competitive representation quality without an adversarial critic. Direct regression to the same assigned ImageNet-1K targets does not reproduce its accuracy under the matched protocol (Appendix E.1, Table 8).

4.2 Throughput and GPU Memory

Table 3 compares native DM and FBDM under a matched single-GPU protocol on CIFAR-10, CIFAR-100, and STL-10; measurement details are given in Appendix D.4.

Table 3: Matched DM–FBDM training cost on one Tesla V100-SXM2-16GB. Memory is peak CUDA allocated/reserved GiB. V100-hours are projected from the median seconds per epoch for 1000 epochs.

Dataset	Method	Peak memory (GiB)	Seconds/epoch	V100-h/1000 ep.	Speedup vs. DM
CIFAR-10	DM	4.78/5.14	75.61	21.00	**1.83×**
	FBDM	5.28/5.82	41.35	11.49	
CIFAR-100	DM	4.78/5.14	76.37	21.21	**1.69×**
	FBDM	5.28/5.82	45.27	12.58	
STL-10	DM	7.06/8.17	173.35	48.15	**1.48×**
	FBDM	7.56/8.94	117.39	32.61	

FBDM is 1.83×, 1.69×, and 1.48× faster per epoch on CIFAR-10, CIFAR-100, and STL-10, saving 9.52, 8.64, and 15.55 projected V100-hours per 1000 epochs. It uses about 0.50 GiB more allocated and 0.67–0.77 GiB more reserved memory; absolute times remain hardware dependent.

5 Theoretical Analysis

We analyze FBDM at the population level to connect its training objective to downstream classification. Pretraining and classification use the same distribution $(\mathbf{X}, Y) \sim P$, with labels hidden during pretraining, $Y \in [K]$, and $p_k = P(Y = k) > 0$; write $p_{\min} = \min_k p_k$. We treat the normalized map in (1) as a single encoder $f : \mathcal{X} \to \mathbb{S}^{d^\star - 1}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ contains images and their views.

5.1 Questions Motivating the Analysis

The velocity field is used only during pretraining and discarded downstream. This raises three linked questions:

1. Why can a velocity field used only during pretraining improve the encoder representation after the field is discarded?
2. Under what conditions does label-free matching to separated reference components induce semantic class separation rather than an arbitrary partition?
3. How do the resulting between-class separation and alignment-controlled within-class variation yield a small downstream misclassification rate?

The proof, detailed in Appendix C.3, follows four steps. First, flow regularity converts the label-free FM risk into a bound on the terminal W_1 discrepancy (Lemma 7). Second, transport identifiability turns this marginal bound into class–component leakage control, placing terminal class means near distinct reference-component means (Lemma 8 and (79)). Third, ODE stability transfers this terminal separation back to the encoder space ((91)). Finally, augmentation and alignment control within-class variation (Lemmas 3 and 4); together with the recovered separation, this yields the downstream classification bound.

5.2 Population Objective and Assumptions

The augmentation kernel $\mathcal{A}$ is fixed. Given $\mathbf{X}$, let $\widetilde{\mathbf{X}}_1$ and $\widetilde{\mathbf{X}}_2$ be conditionally independent draws from $\mathcal{A}(\cdot \mid \mathbf{X})$. Denote the law of one view by $P_{\mathcal{A}}$ and their joint law by $\Pi^+_{\mathcal{A}} = \mathrm{Law}(\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2)$. The reference is also fixed: we consider $K' = K$ and $\nu_\varepsilon = \sum_k w_k Q^\varepsilon_k$, with $w_k > 0$ and $\sum_k w_k = 1$. No equal class priors or equality $p_k = w_k$ is required. Set $\mathcal{S}^\varepsilon_k = \mathrm{supp}(Q^\varepsilon_k)$ and

$$\Delta_{\mathcal{C}} = \min_{k \neq \ell} \|\mathbf{c}_k - \mathbf{c}_\ell\|_2, \qquad r_\varepsilon = \max_k \sup_{\mathbf{r} \in \mathcal{S}^\varepsilon_k} \|\mathbf{r} - \mathbf{c}_k\|_2, \qquad \Delta_{\mathcal{C}} > 2r_\varepsilon. \tag{19}$$

Thus the reference supports are disjoint. Recall the representation–endpoint coupling $\Pi_f = \mathrm{Law}(\mathbf{Z}_0, \mathbf{R})$ from Section 3.3. For the population analysis, its source marginal is the encoded augmentation law $\mathrm{Law}(\mathbf{Z}_0) = f_{\#}P_{\mathcal{A}}$, and its endpoint marginal is $\mathrm{Law}(\mathbf{R}) = \nu_\varepsilon$. The latent label Y remains on the underlying population space but is never used during pretraining. With paths (10), targets (11), and independent $T \sim q_T$, define

$$\mathcal{R}_{\mathrm{FM}} = \frac{1}{d^\star}\mathbb{E}_{\Pi_f,T}[\|\boldsymbol{v}_\phi(\mathbf{Z}_T, T) - \mathbf{U}_T\|_2^2], \tag{20}$$

$$\mathcal{R}_{\mathrm{align}} = \frac{1}{d^\star}\mathbb{E}_{\Pi_{\mathcal{A}}^+}[\|f(\widetilde{\mathbf{X}}_1) - f(\widetilde{\mathbf{X}}_2)\|_2^2], \tag{21}$$

$$\mathcal{R}_{\mathrm{FBDM}}(f,\phi) = \mathcal{R}_{\mathrm{FM}}(f,\phi) + \lambda\mathcal{R}_{\mathrm{align}}(f), \qquad \lambda > 0. \tag{22}$$

See (53) for $\Pi_{\mathcal{A}}^+$ and Table 4 for the analysis notation.

Assumption 1 ((σ,δ)-augmentation and encoder regularity)**.** The kernel $\mathcal{A}$ samples uniformly from a fixed finite family $\mathcal{T} = \{A_1, \ldots, A_{N_{\mathrm{aug}}}\}$ of measurable maps $\mathcal{X} \to \mathcal{X}$, including the identity. For $P_k = \mathrm{Law}(\mathbf{X} \mid Y = k)$, each class has a measurable core $C_k^\circ \subseteq \mathrm{supp}(P_k)$ such that

$$P_k(C_k^\circ) \geq \sigma, \qquad \sup_{\mathbf{x},\mathbf{x}' \in C_k^\circ} \min_{A,A' \in \mathcal{T}} \|A(\mathbf{x}) - A'(\mathbf{x}')\|_2 \leq \delta, \tag{23}$$

where $\sigma \in (0,1]$, $\delta \geq 0$, and f is L_f-Lipschitz.

Here σ measures class-core coverage, while δ measures how closely suitable views connect same-class images (Huang et al., 2023; Duan et al., 2025; Jiao et al., 2026).

Assumption 2 (Uniform flow regularity and time coverage)**.** The field $\boldsymbol{v}_\phi : \mathbb{R}^{d^\star} \times [0,1] \to \mathbb{R}^{d^\star}$ is measurable in t, with

$$\|\boldsymbol{v}_\phi(\mathbf{z},t) - \boldsymbol{v}_\phi(\mathbf{z}',t)\|_2 \leq \ell_\phi(t)\|\mathbf{z} - \mathbf{z}'\|_2, \qquad \int_0^1 \ell_\phi(t)\,\mathrm{d}t \leq L_v < \infty. \tag{24}$$

Here $\ell_\phi \geq 0$ and L_v is fixed over the admissible fields. Also $\int_0^1 \|\boldsymbol{v}_\phi(\mathbf{0},t)\|_2\,\mathrm{d}t < \infty$. Conditional paths are absolutely continuous, with $\mathbb{E}\int_0^1 \|\mathbf{U}_t\|_2^2\,\mathrm{d}t < \infty$ and $\mathbb{E}\|\mathbf{U}_T\|_2^2 < \infty$. The fixed time density satisfies $q_T \geq q_{\min} > 0$ almost everywhere.

For (14), $q_T(t) = \dot{s}(t)$. The schedules used in our experiments satisfy $q_{\min} > 0$; their explicit forms and parameters are listed in Appendix D.3, Table 6.

Let Φ_t solve

$$\frac{\mathrm{d}\Phi_t(\mathbf{z})}{\mathrm{d}t} = \boldsymbol{v}_\phi(\Phi_t(\mathbf{z}),t), \quad \Phi_0(\mathbf{z}) = \mathbf{z}, \quad \widehat{\mathbf{Z}}_t := \Phi_t(\mathbf{Z}_0), \quad \widehat{p}_t := \mathrm{Law}(\widehat{\mathbf{Z}}_t), \quad t \in [0,1]. \tag{25}$$

This learned terminal law need not equal the analytic endpoint law $\mathrm{Law}(\mathbf{R}) = \nu_\varepsilon$. For probability laws μ and ν on $\mathbb{R}^{d^\star}$ with finite first moments, let $\Pi(\mu,\nu)$ denote their set of couplings and define

$$W_1(\mu,\nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \int \|\mathbf{z} - \mathbf{r}\|_2\,\gamma(\mathrm{d}\mathbf{z}, \mathrm{d}\mathbf{r}). \tag{26}$$

Although W_1 does not appear in the training objective, Lemma 7 shows that the FM risk controls this terminal distributional mismatch:

$$W_1(\widehat{p}_1, \nu_\varepsilon) \leq e^{L_v}\sqrt{\frac{d^\star}{q_{\min}}\mathcal{R}_{\mathrm{FM}}}. \tag{27}$$

Semantic identifiability. The label-free bound (27) controls the overall marginal transport but cannot by itself identify classes. Following Jiao et al. (2026), we therefore impose a quantitative condition on the optimal terminal–reference coupling. Let $\Gamma^\star \in \Pi(\widehat{p}_1, \nu_\varepsilon)$ attain the infimum in (26); thus it is the joint law of $(\widehat{\mathbf{Z}}_1, \mathbf{R}^\star)$ only. Labels enter the analysis afterward. Recall $\mathcal{S}_j^\varepsilon = \mathrm{supp}(Q_j^\varepsilon)$ and define

$$\pi_k(\widehat{\mathbf{z}}) := P(Y = k \mid \widehat{\mathbf{Z}}_1 = \widehat{\mathbf{z}}), \qquad q_{kj}^\star := \int \pi_k(\widehat{\mathbf{z}})\mathbf{1}\{\mathbf{r} \in \mathcal{S}_j^\varepsilon\}\,\Gamma^\star(\mathrm{d}\widehat{\mathbf{z}}, \mathrm{d}\mathbf{r}). \tag{28}$$

Let S_K denote the set of permutations of $[K]$ and define

$$\tau^\star \in \arg\max_{\tau \in S_K} \sum_{k=1}^{K} q^\star_{k,\tau(k)}. \tag{29}$$

Then

$$q^\star_k = q^\star_{k,\tau^\star(k)}, \qquad \mathrm{Leak}(\Gamma^\star) = \max_k \{p_k + w_{\tau^\star(k)} - 2q^\star_k\}. \tag{30}$$

Each summand counts outgoing and incoming mass for a class–component pair. This analytical coupling is distinct from the training-induced representation–endpoint coupling Π_f.

Assumption 3 (Transport identifiability)**.** There is such an optimal coupling with

$$\mathrm{Leak}(\Gamma^\star) \le C_{\mathrm{id}} W_1(\widehat{p}_1, \nu_\varepsilon), \tag{31}$$

where $C_{\mathrm{id}} > 0$ is a constant.

Appendix C.4 gives a sufficient soft geometric condition; neither labels nor this analytical condition are used by the training algorithm. Substituting the flow-matching guarantee (27) into the identifiability condition (31) yields

$$\mathrm{Leak}(\Gamma^\star) \le C_{\mathrm{id}} e^{L_v} \sqrt{\frac{d^\star}{q_{\min}} \mathcal{R}_{\mathrm{FM}}}. \tag{32}$$

The remaining argument turns this small leakage into separated terminal class means, transfers that separation back through the flow, and compares it with the encoder's within-class radius.

5.3 A Population Misclassification Guarantee

For each k, let $\mathbf{X} \sim P_k$ and $A \sim \mathrm{Unif}(\mathcal{T})$ be independent. Define the augmented class means and nearest-centroid classifier by

$$\mathbf{m}_k = \mathbb{E}[f(A(\mathbf{X}))], \qquad G_f(\mathbf{x}) = \arg\min_{k \in [K]} \|f(\mathbf{x}) - \mathbf{m}_k\|_2^2, \tag{33}$$

with fixed tie-breaking, and $\mathrm{Err}(G_f) = P\{G_f(\mathbf{X}) \neq Y\}$. Its scores $2\mathbf{m}_k^\top f(\mathbf{x}) - \|\mathbf{m}_k\|_2^2$ are affine in the initial representation; classification does not run the ODE.

Theorem 1 (Population FBDM loss controls classification error)**.** *Suppose Assumptions 1–3 hold. Fix a view-stability tolerance $\varepsilon_{\mathrm{al}} > 0$ such that*

$$e^{-L_v}(\Delta_{\mathcal{C}} - 2r_\varepsilon) > 4L_f\delta + 8\varepsilon_{\mathrm{al}} + 12(1-\sigma). \tag{34}$$

There is an explicit threshold $r_0 > 0$, depending only on these fixed parameters and specified in (84), such that $\mathcal{R}_{\mathrm{FBDM}}(f,\phi) \le r_0$ implies

$$\mathrm{Err}(G_f) \le (1-\sigma) + C_{\mathrm{err}}\sqrt{\mathcal{R}_{\mathrm{FBDM}}(f,\phi)}, \qquad C_{\mathrm{err}} = \frac{N_{\mathrm{aug}}}{\varepsilon_{\mathrm{al}}}\sqrt{\frac{d^\star}{\lambda}}. \tag{35}$$

The residual $1-\sigma$ accounts for images outside class cores. The proof in Appendix C.3 uses FM regression to control terminal Wasserstein error, identifiability to recover terminal class separation, and ODE stability to transfer it to the initial means. Alignment and augmentation quality bound the within-class radius. Thus smaller full loss tightens the classification bound in the stated regime. This result concerns the normalized map f with population centroids; it is not a separate guarantee for the discarded-projector backbone probes. The small-loss condition is sufficient, not an optimization-convergence claim.

6 Conclusion

FBDM combines explicit target geometry with non-adversarial velocity regression on spherical paths. It separates reference granularity K' from flow-space dimension $d^\star$, discards its transport components downstream, and is substantially faster than critic-based DM under matched measurement. Results across benchmarks ranging from CIFAR to ImageNet support explicit geometric design as a practical principle for self-supervised representation learning. Under the stated population assumptions and small-loss condition, our analysis also bounds downstream misclassification from the initial representation in terms of the full FBDM loss and augmentation quality. This explains how a velocity field discarded after pretraining can help shape a useful representation.

Reproducibility Statement

The appendix gives the complete objective, assignment procedure, clock, augmentation and optimization settings, initialization, evaluation rules, and the distinction between online screening and fixed evaluation. The matched training-cost comparison follows the DM implementation cited in Section 4. Anonymized FBDM code is available at https://anonymous.4open.science/r/FBDM-96EC/.

AI Use Statement

The authors developed the research roadmap, initial FBDM idea, and initial algorithmic implementation. Throughout the project, they set the conceptual direction, proposed and evaluated experimental directions, identified the central weakness to address in the theoretical analysis, and specified the goal of relating the pretraining objective to downstream misclassification.

Within this author-defined framework, OpenAI's ChatGPT and Codex served as substantial research-assistance tools. They helped develop and implement the theoretical details, including the population-level analysis, intermediate mathematical arguments and bounds, and proof drafts toward the authors' specified goal. They also assisted with algorithmic refinement, experimental design, implementation and review of training and evaluation code, debugging, interpretation of intermediate experimental results, literature search, manuscript drafting and revision, figure and table preparation, and LaTeX.

The authors critically examined AI-assisted suggestions, derivations, and experimental proposals through iterative discussion; chose which ideas and experiments to pursue; verified the implementations, proofs, citations, and experimental outputs; and made all final scientific and presentation decisions. The authors take full responsibility for the accuracy, integrity, and conclusions of the method, theoretical analysis, experiments, citations, and reported results.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223, 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Changxiao Cai and Gen Li. Minimax optimality of the probability flow ODE for diffusion models, 2025. URL https://arxiv.org/abs/2503.09583.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis, 2026. URL https://arxiv.org/abs/2603.06507.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g7ohDlTITL.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Chenguang Duan, Yuling Jiao, Huazhen Lin, Wensen Ma, and Jerry Yang. Adv-SSL: Adversarial self-supervised representation learning with theoretical guarantees. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4097. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b216ce49d808157ba14adec9453983a4-Abstract-Conference.html.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, 2021.

Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves almost minimax optimal convergence. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/44a427b77e9727fcec560e2f8d6925e0-Abstract-Conference.html.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow ODEs of diffusion models in wasserstein distances. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1009–1017, 2025. URL https://proceedings.mlr.press/v258/gao25c.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5000–5011, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/27debb435021eb68b3965290b5e24c49-Abstract.html.

Jeff Z. HaoChen, Colin Wei, Ananya Kumar, and Tengyu Ma. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26889–26902, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ac112e8ffc4e5b9ece32070440a8ca43-Abstract-Conference.html.

Junlin He, Jinxiao Du, and Wei Ma. Preventing dimensional collapse in self-supervised learning via orthogonality regularization. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3028. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad7922fd4650f8aba5d8b067e622ca84-Abstract-Conference.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9598–9608, 2021.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ODE for score-based generative models. *IEEE Transactions on Information Theory*, 71(6): 4581–4601, 2025. doi: 10.1109/TIT.2025.3557050. URL https://doi.org/10.1109/TIT.2025.3557050.

Hanxun Huang, Ricardo J. G. B. Campello, Sarah Monazam Erfani, Xingjun Ma, Michael E. Houle, and James Bailey. LDReg: Local dimensionality regularized self-supervised learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oZyAqjAjJW.

Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. Towards the generalization of contrastive self-supervised learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XDJwuEYHhme.

Yuling Jiao, Wensen Ma, Defeng Sun, Hansheng Wang, and Yang Wang. Bringing generative learning to representation learning: Self-supervised transfer learning as distribution matching, 2026. URL https://arxiv.org/abs/2502.14424.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow ODEs of diffusion models, 2024. URL https://arxiv.org/abs/2408.02320.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

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

Achleshwar Luthra, Priyadarsi Mishra, and Tomer Galanti. On the alignment between supervised and self-supervised contrastive learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fa76985f05e0a25c66528308dda33de0-Abstract-Conference.html.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637, 2019. URL https://proceedings.mlr.press/v97/saunshi19a.html.

Oriane Siméoni et al. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Jiaqi Tang and Yuling Yan. Adaptivity and convergence of probability flow ODEs in diffusion generative models, 2025. URL https://arxiv.org/abs/2501.18863.

Kosuke Ukita and Tsuyoshi Okita. High-performance self-supervised learning by joint training of flow matching. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4492–4500. PMLR, 2026. URL https://proceedings.mlr.press/v300/ukita26a.html.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 2020.

Lloyd R. Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information Theory*, 20(3):397–399, 1974. doi: 10.1109/TIT.1974.1055219. URL https://doi.org/10.1109/TIT.1974.1055219.

Xi Weng, Lei Huang, Lei Zhao, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. An investigation into whitening loss for self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2157. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c057cb81b8d3c67093427bf1c16a4e9f-Abstract-Conference.html.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d9e42b4d7163931f3689d6d6fbaa11d0-Abstract-Conference.html.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 2021.

Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-CL: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RAW9tCdVxLj.

Yipeng Zhang, Hafez Ghaemi, Jungyoon Lee, Shahab Bakhtiari, Eilif B. Muller, and Laurent Charlin. Self-supervised learning from structural invariance. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d26d1923ce22d5afc3ee0db15ef4d900-Abstract-Conference.html.

A Reference Geometry

This appendix supports the reference construction in Section 3.1. Appendix A.1 establishes the angular benchmark provided by ETFs; Appendix A.2 verifies the exact simplex ETF encoded by $\mathbf{G}_\Delta$ and gives its spectral construction; Appendix A.3 justifies the lower-dimensional approximation.

A.1 ETF Angular Benchmark

We first ask how far apart K' unit centers can be placed in $\mathbb{R}^{d^\star}$. The following definition and proposition formalize the ETF benchmark invoked in Section 3.1.

Definition 1 (Equiangular tight frame)**.** Let $K' \geq d^\star \geq 2$. A collection $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{K'} \subset \mathbb{S}^{d^\star-1}$ is an equiangular tight frame (ETF) if

$$\sum_{k=1}^{K'} \mathbf{c}_k \mathbf{c}_k^\top = \frac{K'}{d^\star} \mathbf{I}_{d^\star}, \qquad |\mathbf{c}_k^\top \mathbf{c}_\ell| = \alpha \quad (k \neq \ell) \tag{36}$$

for a constant $\alpha \geq 0$. These conditions express tightness and equiangularity, respectively.

For $K' \geq d^\star$, define

$$\mu(\mathcal{C}) := \max_{k \neq \ell} |\mathbf{c}_k^\top \mathbf{c}_\ell|, \qquad \theta_{\min}^{\pm}(\mathcal{C}) := \arccos(\mu(\mathcal{C})). \tag{37}$$

Proposition 1 (ETF angular benchmark)**.** *Let $K' \geq d^\star \geq 2$. Every collection $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{K'} \subset \mathbb{S}^{d^\star-1}$ satisfies*

$$\mu(\mathcal{C}) \geq \mu_{\mathrm{W}} := \sqrt{\frac{K' - d^\star}{d^\star(K'-1)}}, \qquad \theta_{\min}^{\pm}(\mathcal{C}) \leq \arccos(\mu_{\mathrm{W}}). \tag{38}$$

Equality holds if and only if $\mathcal{C}$ is an ETF in the sense of Definition 1.

Equivalently, every placement of K' centers contains at least one pair whose sign-invariant angle is at most $\arccos(\mu_{\mathrm{W}})$; when an ETF exists, it makes this worst-pair angle as large as possible (Welch, 1974).

Proof of Proposition 1. Assume $K' \geq d^\star \geq 2$. Let $\mathbf{C} \in \mathbb{R}^{K' \times d^\star}$ have kth row $\mathbf{c}_k^\top$, and let $\mathbf{G} = \mathbf{C}\mathbf{C}^\top$. With $\mu = \max_{k \neq \ell} |\mathbf{c}_k^\top \mathbf{c}_\ell|$,

$$\|\mathbf{G}\|_F^2 = K' + \sum_{k \neq \ell} |\mathbf{c}_k^\top \mathbf{c}_\ell|^2 \leq K' + K'(K'-1)\mu^2. \tag{39}$$

Thus, a lower bound on μ follows from a lower bound on $\|\mathbf{G}\|_F^2$. Let $r = \operatorname{rank}(\mathbf{G}) \leq d^\star$ and let $\lambda_1(\mathbf{G}), \ldots, \lambda_r(\mathbf{G}) > 0$ be its nonzero eigenvalues. Since $\mathbf{G}$ is positive semidefinite and the centers have unit norm, $\operatorname{tr}(\mathbf{G}) = \sum_{i=1}^r \lambda_i(\mathbf{G}) = K'$. Cauchy–Schwarz therefore gives

$$\begin{aligned} \|\mathbf{G}\|_F^2 = \sum_{i=1}^r \lambda_i(\mathbf{G})^2 &\geq \frac{1}{r} \left\{ \sum_{i=1}^r \lambda_i(\mathbf{G}) \right\}^2 \\ &= \frac{\operatorname{tr}(\mathbf{G})^2}{r} \geq \frac{\operatorname{tr}(\mathbf{G})^2}{d^\star} = \frac{K'^2}{d^\star}. \end{aligned} \tag{40}$$

The first inequality is Cauchy–Schwarz, and the second uses $r \leq d^\star$. Substituting this bound into the preceding inequality yields

$$\mu^2 \geq \frac{K' - d^\star}{d^\star(K'-1)}, \tag{41}$$

which is (38). Equality requires equality in both preceding steps. Equality in the first bound means that all off-diagonal inner products have the same magnitude, which is equiangularity. In the eigenvalue bound, equality in Cauchy–Schwarz requires $\lambda_1(\mathbf{G}) = \cdots = \lambda_r(\mathbf{G})$, while equality in the step $r \leq d^\star$ requires $r = d^\star$. Together these conditions are equivalent to $\mathbf{C}^\top \mathbf{C} = (K'/d^\star)\mathbf{I}_{d^\star}$, which is tightness. The converse follows by substituting the ETF identities into the two bounds. Because $\arccos$ is decreasing on $[0, 1]$, $\theta_{\min}^{\pm}(\mathcal{C}) = \arccos(\mu(\mathcal{C}))$ gives the angular inequality in (38), with equality under exactly the same ETF conditions. □

A.2 Simplex Gram Matrix and Spectral Construction

The construction in Section 3.1 starts from the following goal: realize K' unit vectors in $\mathbb{R}^{K'-1}$ that are simultaneously equiangular and tight. Rather than choosing their coordinates directly, we first encode the desired geometry in the ideal Gram matrix $\mathbf{G}_\Delta$ from (2), whose entries are

$$(\mathbf{G}_\Delta)_{k\ell} = \begin{cases} 1, & k = \ell, \\ -1/(K'-1), & k \neq \ell. \end{cases}$$

The diagonal directly prescribes unit row norms, while the off-diagonal entries prescribe the common inner product $-1/(K'-1)$.

The matrix $\mathbf{G}_\Delta$ is positive semidefinite with $\operatorname{rank}(\mathbf{G}_\Delta) = K'-1$. Its $K'-1$ positive eigenvalues are all $K'/(K'-1)$, and its remaining eigenvalue is zero. Equivalently, its full spectral decomposition is

$$\mathbf{G}_\Delta = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \qquad \mathbf{U} \in \mathbb{R}^{K'\times K'}, \quad \mathbf{\Lambda} = \operatorname{diag}\left(\frac{K'}{K'-1}\mathbf{I}_{K'-1}, 0\right) \in \mathbb{R}^{K'\times K'}.$$

Thus the Gram realization theorem guarantees K' vectors in $\mathbb{R}^{K'-1}$ with this Gram matrix.

For an explicit realization, let $\mathbf{U}_+ \in \mathbb{R}^{K'\times(K'-1)}$ and $\mathbf{\Lambda}_+ \in \mathbb{R}^{(K'-1)\times(K'-1)}$ denote the eigenvectors and diagonal block associated with the positive eigenvalues. The subscript $+$ always refers to this positive-eigenvalue block. Since the omitted spectral term is $0\,\mathbf{u}_0\mathbf{u}_0^\top = \mathbf{0}$, the thin factorization remains exact:

$$\mathbf{G}_\Delta = \mathbf{U}_+\mathbf{\Lambda}_+\mathbf{U}_+^\top, \qquad \mathbf{\Lambda}_+ = \frac{K'}{K'-1}\mathbf{I}_{K'-1}.$$

Set $\mathbf{C}_\Delta = \mathbf{U}_+\mathbf{\Lambda}_+^{1/2} \in \mathbb{R}^{K'\times(K'-1)}$. Then

$$\mathbf{C}_\Delta\mathbf{C}_\Delta^\top = \mathbf{G}_\Delta, \qquad \mathbf{C}_\Delta^\top\mathbf{C}_\Delta = \frac{K'}{K'-1}\mathbf{I}_{K'-1}.$$

The first identity shows that the rows of $\mathbf{C}_\Delta$ have unit norm and pairwise inner product $-1/(K'-1)$; the second establishes tightness. Therefore these rows form the exact simplex ETF of K' vectors in $\mathbb{R}^{K'-1}$ from Definition 1.

Only now do we reduce the dimension. For $2 \le d^\star < K'-1$, retain $d^\star$ directions from the positive eigenspace and their eigenvalues:

$$\mathbf{U}_{d^\star} \in \mathbb{R}^{K'\times d^\star}, \qquad \mathbf{\Lambda}_{d^\star} = \frac{K'}{K'-1}\mathbf{I}_{d^\star} \in \mathbb{R}^{d^\star\times d^\star}.$$

Equation (3) forms $\mathbf{Y}$ from these matrices. Lemma 2 shows that $\mathbf{Y}\mathbf{Y}^\top$ is an optimal rank-$d^\star$ approximation of the ideal Gram geometry. At $d^\star = K'-1$, retaining the entire positive eigenspace recovers the exact simplex ETF.

Algorithm 2 summarizes the complete construction.

Algorithm 2 ETF-inspired simplex-spectral center construction.

Require: Number of centers K' and matching dimension $2 \le d^\star \le K'-1$
Ensure: Unit centers $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{K'} \subset \mathbb{S}^{d^\star-1}$

1: Form $\mathbf{G}_\Delta \leftarrow \frac{K'}{K'-1}\left(\mathbf{I}_{K'} - \frac{\mathbf{1}\mathbf{1}^\top}{K'}\right) \in \mathbb{R}^{K'\times K'}$
2: Compute its nonzero eigenspace, with eigenvalue $K'/(K'-1)$
3: Select $d^\star$ orthonormal vectors from this eigenspace as $\mathbf{U}_{d^\star} \in \mathbb{R}^{K'\times d^\star}$
4: Set $\mathbf{\Lambda}_{d^\star} \leftarrow \frac{K'}{K'-1}\mathbf{I}_{d^\star}$ and $\mathbf{Y} \leftarrow \mathbf{U}_{d^\star}\mathbf{\Lambda}_{d^\star}^{1/2}$, with rows $\mathbf{y}_k^\top$
5: **for** $k = 1, \dots, K'$ **do**
6: $\quad \mathbf{c}_k \leftarrow \mathbf{y}_k/\|\mathbf{y}_k\|_2$
7: **return** $\mathcal{C}$

A.3 Spectral Approximation

We finally justify the approximation claim following (3).

Using the eigenpairs defined above, the projected coordinate matrix is

$$\mathbf{Y} = \mathbf{U}_{d^\star} \mathbf{\Lambda}_{d^\star}^{1/2} \in \mathbb{R}^{K' \times d^\star}, \qquad \mathbf{Y}_{k,:} = \mathbf{y}_k^\top .$$

The following result concerns these coordinates before row normalization.

Lemma 2 (Optimal preservation of pairwise simplex geometry)**.** *Let $K' > d^\star \geq 2$, and let $\widetilde{\mathbf{y}}_k^\top$ denote the kth row of $\widetilde{\mathbf{Y}}$. Among all configurations of K' points in $\mathbb{R}^{d^\star}$, the spectral coordinates in (3) minimize the total squared distortion of the ideal simplex pairwise inner products:*

$$\mathbf{Y} \in \arg \min_{\widetilde{\mathbf{Y}} \in \mathbb{R}^{K' \times d^\star}} \left\| \mathbf{G}_\Delta - \widetilde{\mathbf{Y}} \widetilde{\mathbf{Y}}^\top \right\|_F^2 , \tag{42}$$

$$\left\| \mathbf{G}_\Delta - \widetilde{\mathbf{Y}} \widetilde{\mathbf{Y}}^\top \right\|_F^2 = \sum_{k,\ell=1}^{K'} \left\{ (\mathbf{G}_\Delta)_{k\ell} - \widetilde{\mathbf{y}}_k^\top \widetilde{\mathbf{y}}_\ell \right\}^2 .$$

The first line is the classical Eckart–Young low-rank approximation result (Eckart & Young, 1936); the second line makes its geometry explicit. Dividing by K'^2 shows that spectral projection minimizes the mean squared error over all corresponding pairwise inner products before row normalization. Row normalization then places the projected coordinates on $\mathbb{S}^{d^\star - 1}$ without changing their directions or pairwise angles. We therefore call (3) an *ETF-inspired simplex-spectral reference*, not an exact ETF.

B Flow-Matching Derivations

B.1 Spherical Path and Target Velocity

Fix non-antipodal $\mathbf{z}, \mathbf{r} \in \mathbb{S}^{d^\star - 1}$ and let $\omega = \arccos(\mathbf{z}^\top \mathbf{r})$. Since $\|\mathbf{z}\|_2 = 1$, $\mathbf{z}^\top \mathbf{r}$ is the scalar projection of $\mathbf{r}$ onto $\mathbf{z}$. Thus $\mathbf{r}$ admits the orthogonal decomposition

$$\mathbf{r} = \underbrace{(\mathbf{z}^\top \mathbf{r})\mathbf{z}}_{\text{normal component}} + \underbrace{\left[\mathbf{r} - (\mathbf{z}^\top \mathbf{r})\mathbf{z}\right]}_{\text{tangent component}} = \cos(\omega)\mathbf{z} + \left[\mathbf{r} - \cos(\omega)\mathbf{z}\right] . \tag{43}$$

The two components are orthogonal. For $0 < \omega < \pi$, the second has norm $\sin(\omega)$, so normalizing it gives

$$\mathbf{u}_\perp := \frac{\mathbf{r} - \cos(\omega)\mathbf{z}}{\sin(\omega)}. \tag{44}$$

Then $\{\mathbf{z}, \mathbf{u}_\perp\}$ is an orthonormal basis of the plane spanned by $\mathbf{z}$ and $\mathbf{r}$. The intersection of this plane with the unit sphere is the great circle containing both endpoints. Its arc from $\mathbf{z}$ to $\mathbf{r}$ is therefore

$$\boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(s) = \cos(s\omega)\mathbf{z} + \sin(s\omega)\mathbf{u}_\perp, \qquad s \in [0, 1]. \tag{45}$$

Substituting $\mathbf{u}_\perp$ and using $\sin((1-s)\omega) = \sin(\omega)\cos(s\omega) - \cos(\omega)\sin(s\omega)$ gives

$$\boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(s) = \left[\cos(s\omega) - \frac{\cos(\omega)\sin(s\omega)}{\sin(\omega)}\right]\mathbf{z} + \frac{\sin(s\omega)}{\sin(\omega)}\mathbf{r} \tag{46}$$

$$= \frac{\sin((1-s)\omega)}{\sin(\omega)}\mathbf{z} + \frac{\sin(s\omega)}{\sin(\omega)}\mathbf{r}, \tag{47}$$

which is (9). Moreover,

$$\partial_s \boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(s) = \omega \left[-\sin(s\omega)\mathbf{z} + \cos(s\omega)\mathbf{u}_\perp\right], \tag{48}$$

Equivalently, the expanded derivative is

$$\partial_s \boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(s) = -\frac{\omega \cos((1-s)\omega)}{\sin \omega}\mathbf{z} + \frac{\omega \cos(s\omega)}{\sin \omega}\mathbf{r}. \tag{49}$$

Orthogonality yields

$$\|\boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(s)\|_2 = 1, \qquad \boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(s)^\top \partial_s \boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(s) = 0, \qquad \|\partial_s \boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(s)\|_2 = \omega. \tag{50}$$

The endpoints are $\boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(0) = \mathbf{z}$ and $\boldsymbol{\gamma}_{\mathbf{z},\mathbf{r}}(1) = \mathbf{r}$, and the curve length is $\int_0^1 \omega \, \mathrm{d}s = \omega$, equal to the spherical distance $\arccos(\mathbf{z}^\top \mathbf{r})$. Hence it is the unique shortest geodesic for $0 < \omega < \pi$. At $\omega = 0$ the same formula gives the constant path by continuity; the antipodal case $\omega = \pi$, where the shortest geodesic is not unique, is excluded in Section 3.

C Theoretical Analysis

C.1 Key Notation and Proof Sketch

Table 4: Key notation for the flow and classification analysis. The index k denotes a semantic class, and j denotes a reference component. All representation vectors lie in $\mathbb{R}^{d^\star}$.

Symbol	Meaning / definition	First defined
$\mathbf{Z}_0$	Encoder representation $f(\widetilde{\mathbf{X}})$ before ODE evolution; used for downstream classification.	Section 3.3
$\mathbf{R}$	Assigned reference target; the prescribed endpoint of the analytic path.	Section 3.3
$\mathbf{Z}_t$	State on the prescribed analytic path, with $\mathbf{Z}_1 = \mathbf{R}$.	(10)
$\mathbf{U}_t$	Conditional target velocity $\mathrm{d}\mathbf{Z}_t/\mathrm{d}t$.	(11)
Φ_t	Learned ODE map satisfying $\mathrm{d}\Phi_t(\mathbf{z})/\mathrm{d}t = \boldsymbol{v}_\phi(\Phi_t(\mathbf{z}), t)$ and $\Phi_0(\mathbf{z}) = \mathbf{z}$.	(25)
$\widehat{\mathbf{Z}}_t$	Learned ODE state $\Phi_t(\mathbf{Z}_0)$; $\widehat{\mathbf{Z}}_1$ is its terminal representation, not the prescribed target.	(25)
$\mathbf{m}_k$	Class mean before ODE evolution: $\mathbb{E}[\mathbf{Z}_0 \mid Y = k]$.	(33)
$\widehat{\mathbf{m}}_k$	Class mean after learned ODE evolution: $\mathbb{E}[\widehat{\mathbf{Z}}_1 \mid Y = k]$; generally not $\Phi_1(\mathbf{m}_k)$.	(73)
$\mathbf{c}_j$	Geometric center of reference component j.	Section 3.1
$\mathbf{b}_j$	Reference component mean $\int \mathbf{r}\, Q_j^\varepsilon(\mathrm{d}\mathbf{r})$; generally not $\mathbf{c}_j$.	(73)
ν_ε	Fixed reference mixture $\sum_j w_j Q_j^\varepsilon$.	(4); Section 5.2
$\widehat{p}_1$	Learned terminal marginal $\mathrm{Law}(\widehat{\mathbf{Z}}_1)$.	(25)
Π_f	Training representation–target coupling $\mathrm{Law}(\mathbf{Z}_0, \mathbf{R})$; global transport optimality is not required.	Section 3.3
$\Gamma^\star$	Analytical optimal W_1 coupling of $\widehat{p}_1$ and ν_ε: $\mathrm{Law}(\widehat{\mathbf{Z}}_1, \mathbf{R}^\star)$.	Section 5.2

Proof sketch. The proof of Theorem 1 distinguishes the class means before and after the learned ODE, $\mathbf{m}_k$ and $\widehat{\mathbf{m}}_k$, from the reference component mean $\mathbf{b}_j$ and its geometric center $\mathbf{c}_j$. Table 4 records their definitions. The argument proceeds in four steps.

1. **Match terminal class means to reference means.** Lemma 7 bounds $W_1(\widehat{p}_1, \nu_\varepsilon)$ by the square root of the FM risk. Under the leakage condition in Assumption 3, Lemma 8 converts this overall distributional bound into a bound for each terminal class mean. Using $\mathcal{R}_{\mathrm{FM}} \le \mathcal{R}_{\mathrm{FBDM}}$ gives

$$\|\widehat{\mathbf{m}}_k - \mathbf{b}_{\tau^\star(k)}\|_2 \le \frac{(1 + 2C_{\mathrm{id}})e^{L_v}}{p_{\min}} \sqrt{\frac{d^\star}{q_{\min}} \mathcal{R}_{\mathrm{FBDM}}}.$$

   Thus a small FBDM loss makes each ODE-terminal class mean close to the mean of its corresponding reference component.

2. **Obtain separation at the ODE endpoint.** The reference component means themselves are separated: each lies within r_ε of its geometric center, so

$$\|\mathbf{b}_i - \mathbf{b}_j\|_2 \ge \Delta_{\mathcal{C}} - 2r_\varepsilon > 0, \qquad i \ne j.$$

Since each terminal class mean is close to a different reference component mean, the triangle inequality transfers this separation to $\widehat{\mathbf{m}}_k$ and $\widehat{\mathbf{m}}_\ell$ when the loss-dependent errors in Step 1 are sufficiently small.

3. **Recover separation of the encoder class means.** Downstream classification uses the initial encoder representations $\mathbf{Z}_0$, so the class means we actually need to separate are $\mathbf{m}_k$ and $\mathbf{m}_\ell$, not their terminal counterparts. Lemma 7 establishes the forward Lipschitz bound

$$\|\Phi_1(\mathbf{m}_k) - \Phi_1(\mathbf{m}_\ell)\|_2 \le e^{L_v}\|\mathbf{m}_k - \mathbf{m}_\ell\|_2.$$

The ODE can enlarge distances by at most e^{L_v}, so separated evolved means require separated initial means. Here $\Phi_1(\mathbf{m}_k)$ evolves the initial class mean, whereas $\widehat{\mathbf{m}}_k$ averages the individually evolved representations. Equation (86) in Section C.3 gives

$$\|\Phi_1(\mathbf{m}_k) - \widehat{\mathbf{m}}_k\|_2 \le e^{L_v}(h_{\mathrm{aug}} + 4b_{\mathrm{aug}}).$$

Subtracting these discrepancies from the terminal separation and dividing by e^{L_v} yields the initial class-mean distance lower bound (91).

4. **Turn separation and concentration into classification.** Lemma 4 gives the within-class concentration bound (62):

$$\mathbb{E}_{\mathbf{X}\sim P_k,\, A\sim\mathrm{Unif}(\mathcal{T})}[\|f(A(\mathbf{X})) - \mathbf{m}_k\|_2] \le h_{\mathrm{aug}} + 4b_{\mathrm{aug}}.$$

Combining Steps 1–3 gives the initial class-mean distance bound (91). Write $\Delta_f := \min_{k\neq\ell}\|\mathbf{m}_k - \mathbf{m}_\ell\|_2$. Substituting the loss bound from Lemma 7 gives

$$\min_{k\neq\ell}\|\mathbf{m}_k - \mathbf{m}_\ell\|_2 \ge e^{-L_v}(\Delta_{\mathcal{C}} - 2r_\varepsilon) - 2h_{\mathrm{aug}} - 8b_{\mathrm{aug}} - \frac{2(1+2C_{\mathrm{id}})}{p_{\min}}\sqrt{\frac{d^\star}{q_{\min}}\mathcal{R}_{\mathrm{FBDM}}}.$$

Thus the encoder representations already have within-class concentration and between-class separation at the ODE's initial position. More precisely, (63) and the theorem's small-loss condition (93) give

$$\|f(\mathbf{x}) - \mathbf{m}_k\|_2 \le h_{\mathrm{aug}} + 2b_{\mathrm{aug}} < \frac{\Delta_f}{2}, \qquad \mathbf{x}\in C_k^\circ \cap \mathcal{V}.$$

The triangle inequality therefore makes each stable class-core representation closer to its own mean than to any other class mean. Only observations outside these cores or with unstable views can be misclassified. Lemma 3 bounds their mass, giving Theorem 1:

$$\mathrm{Err}(G_f) \le (1-\sigma) + C_{\mathrm{err}}\sqrt{\mathcal{R}_{\mathrm{FBDM}}}.$$

The FM risk ensures sufficient between-class separation, while alignment controls within-class concentration and the unstable mass.

Section C.2 fixes the probability laws, and Section C.3 gives the full proof. Sections C.4 and C.5 discuss a sufficient leakage condition and the implemented batchwise coupling, respectively.

C.2 Probability Laws

Let $(\mathbf{X}, Y) \sim P$, with $Y \in [K]$ and $K \ge 2$. We require only

$$p_k := P(Y = k) > 0, \qquad p_{\min} := \min_{k\in[K]} p_k > 0. \tag{51}$$

For measurable $E \subseteq \mathcal{X}$, define the marginal law $P_X(E) := P(\mathbf{X} \in E)$ and let $P_k := \mathrm{Law}(\mathbf{X} \mid Y = k)$. The augmentation kernel $\mathcal{A}$ is fixed independently of the networks. Given $\mathbf{X}$, draw $\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2$ independently from $\mathcal{A}(\cdot \mid \mathbf{X})$. For measurable $E, E_1, E_2 \subseteq \mathcal{X}$, define the augmented-image and positive-pair laws by

$$P_{\mathcal{A}}(E) := \int \mathcal{A}(E \mid \mathbf{x})\, P_X(\mathrm{d}\mathbf{x}), \tag{52}$$

$$\Pi^+_{\mathcal{A}}(E_1 \times E_2) := \int \mathcal{A}(E_1 \mid \mathbf{x})\mathcal{A}(E_2 \mid \mathbf{x})\, P_X(\mathrm{d}\mathbf{x}). \tag{53}$$

Define $P_{\mathcal{A},k}$ by replacing P_X with P_k in (52). The positive-pair law retains the shared-image dependence and is generally not $P_{\mathcal{A}} \otimes P_{\mathcal{A}}$.

For later use, the augmentation distance is

$$d_{\mathcal{T}}(\mathbf{x}, \mathbf{x}') := \min_{A,A' \in \mathcal{T}} \|A(\mathbf{x}) - A'(\mathbf{x}')\|_2. \tag{54}$$

C.3 Proof of Theorem 1

We combine the stable-core argument of Huang et al. (2023); Jiao et al. (2026) with terminal flow matching and a pullback of class-centre separation. The classifier centres are averages of augmented representations, whereas the classified inputs are original images. All constants use the fixed bounds in Section 5.

Lemma 3 (Alignment controls unstable-view probability)**.** *For the tolerance $\varepsilon_{\mathrm{al}}$ in Theorem 1, define*

$$\begin{aligned} \mathcal{V} &:= \left\{ \mathbf{x} \in \mathrm{supp}(P_X) : \max_{A,A' \in \mathcal{T}} \|f(A(\mathbf{x})) - f(A'(\mathbf{x}))\|_2 \le \varepsilon_{\mathrm{al}} \right\}, \\ \mathcal{U} &:= P_X(\mathcal{V}^c). \end{aligned} \tag{55}$$

Under Assumption 1,

$$\mathcal{U} \le \frac{N_{\mathrm{aug}}}{\varepsilon_{\mathrm{al}}} \sqrt{d^\star \mathcal{R}_{\mathrm{align}}} \le C_{\mathrm{err}} \sqrt{\mathcal{R}_{\mathrm{FBDM}}}. \tag{56}$$

Proof. For each fixed $\mathbf{x}$, the maximum of the pairwise distances is at most their squared-sum norm. Uniform sampling from $\mathcal{T}$ gives

$$\begin{aligned} \max_{A,A' \in \mathcal{T}} \|f(A(\mathbf{x})) - f(A'(\mathbf{x}))\|_2 &\le \left[\sum_{i,j=1}^{N_{\mathrm{aug}}} \|f(A_i(\mathbf{x})) - f(A_j(\mathbf{x}))\|_2^2 \right]^{1/2} \\ &= N_{\mathrm{aug}} \left[\mathbb{E}_{A,A'} \|f(A(\mathbf{x})) - f(A'(\mathbf{x}))\|_2^2 \right]^{1/2}. \end{aligned} \tag{57}$$

Markov's inequality first gives

$$\begin{aligned} \mathcal{U} &= P_X\left(\left\{ \mathbf{x} : \max_{A,A' \in \mathcal{T}} \|f(A(\mathbf{x})) - f(A'(\mathbf{x}))\|_2 > \varepsilon_{\mathrm{al}} \right\} \right) \\ &\le \frac{1}{\varepsilon_{\mathrm{al}}} \mathbb{E}_{\mathbf{X} \sim P_X} \left[\max_{A,A' \in \mathcal{T}} \|f(A(\mathbf{X})) - f(A'(\mathbf{X}))\|_2 \right]. \end{aligned} \tag{58}$$

Using the pointwise bound above and then Cauchy–Schwarz,

$$\begin{aligned} &\mathbb{E}_{\mathbf{X} \sim P_X} \left[\max_{A,A' \in \mathcal{T}} \|f(A(\mathbf{X})) - f(A'(\mathbf{X}))\|_2 \right] \\ &\le N_{\mathrm{aug}}\, \mathbb{E}_{\mathbf{X} \sim P_X} \left[\left\{ \mathbb{E}_{A,A'} \|f(A(\mathbf{X})) - f(A'(\mathbf{X}))\|_2^2 \right\}^{1/2} \right] \\ &\le N_{\mathrm{aug}} \left[\mathbb{E}_{\mathbf{X},A,A'} \|f(A(\mathbf{X})) - f(A'(\mathbf{X}))\|_2^2 \right]^{1/2} \\ &= N_{\mathrm{aug}} \sqrt{d^\star \mathcal{R}_{\mathrm{align}}}. \end{aligned} \tag{59}$$

Combining (58) and (59), and using $\lambda \mathcal{R}_{\mathrm{align}} \le \mathcal{R}_{\mathrm{FBDM}}$, yields

$$\mathcal{U} \le \frac{N_{\mathrm{aug}}}{\varepsilon_{\mathrm{al}}} \sqrt{d^\star \mathcal{R}_{\mathrm{align}}} \le \frac{N_{\mathrm{aug}}}{\varepsilon_{\mathrm{al}}} \sqrt{\frac{d^\star}{\lambda}} \sqrt{\mathcal{R}_{\mathrm{FBDM}}} = C_{\mathrm{err}} \sqrt{\mathcal{R}_{\mathrm{FBDM}}}. \tag{60}$$

Thus N_{aug} is precisely the cost of converting the average pairwise alignment error over the finite augmentation family into uniform control over all augmentation pairs. □

The next lemma turns view-level alignment into class-level concentration. The (σ, δ)-augmentation assumption connects same-class images through nearby views, while Lemma 3 controls the images whose views are not aligned. Together with Lipschitz continuity, these facts place most class-k representations near their class mean. This within-class radius is used later to convert the separation pulled back by the learned flow into a nearest-centroid guarantee.

Lemma 4 (Augmentation cores control within-class distances)**.** *Recall from (33) that* $\mathbf{m}_k = \mathbb{E}[f(A(\mathbf{X}))]$ *for independent* $\mathbf{X} \sim P_k$ *and* $A \sim \mathrm{Unif}(\mathcal{T})$*, and set*

$$h_{\mathrm{aug}} := L_f\delta + 2\varepsilon_{\mathrm{al}}, \qquad b_{\mathrm{aug}} := 1 - \sigma + \frac{\mathcal{U}}{p_{\min}}. \tag{61}$$

Under Assumption 1, for every k*,*

$$\mathbb{E}_{\mathbf{X}\sim P_k,\, A\sim\mathrm{Unif}(\mathcal{T})}[\|f(A(\mathbf{X})) - \mathbf{m}_k\|_2] \le h_{\mathrm{aug}} + 4b_{\mathrm{aug}}. \tag{62}$$

Moreover, every original image $\mathbf{x} \in C_k^\circ \cap \mathcal{V}$ *satisfies*

$$\|f(\mathbf{x}) - \mathbf{m}_k\|_2 \le h_{\mathrm{aug}} + 2b_{\mathrm{aug}}. \tag{63}$$

The ideal regime is transparent from these bounds. As $\sigma \to 1$*,* $\delta \to 0$*,* $\varepsilon_{\mathrm{al}} \to 0$*, and* $\mathcal{U} \to 0$*, both* h_{aug} *and* b_{aug} *vanish, so the within-class radius contracts to zero. In the exact limit, (62) gives* $\mathbf{Z}_0 = \mathbf{m}_k$ *almost surely conditional on* $Y = k$*; hence the class-conditional representation has zero variance. Near this limit, the two bounds quantify how imperfect augmentation coverage, augmented distance, and view alignment broaden each class around its mean.*

Proof. The union bound, $P_k(C_k^\circ) \ge \sigma$, and $p_k \ge p_{\min}$ give

$$\begin{aligned} P_k(\{C_k^\circ \cap \mathcal{V}\}^c) &\le P_k(\{C_k^\circ\}^c) + P_k(\mathcal{V}^c) \\ &\le 1 - \sigma + \frac{P(\mathbf{X} \in \mathcal{V}^c, Y = k)}{p_k} \\ &\le 1 - \sigma + \frac{\mathcal{U}}{p_k} \le 1 - \sigma + \frac{\mathcal{U}}{p_{\min}} = b_{\mathrm{aug}}. \end{aligned} \tag{64}$$

For $\mathbf{x}, \mathbf{x}' \in C_k^\circ \cap \mathcal{V}$, choose $A_*, A_*' \in \mathcal{T}$ attaining the minimum in (54). For any $A, A' \in \mathcal{T}$, Lipschitz continuity and view stability give

$$\begin{aligned} \|f(A(\mathbf{x})) - f(A'(\mathbf{x}'))\|_2 &\le \|f(A(\mathbf{x})) - f(A_*(\mathbf{x}))\|_2 + \|f(A_*(\mathbf{x})) - f(A_*'(\mathbf{x}'))\|_2 \\ &\quad + \|f(A_*'(\mathbf{x}')) - f(A'(\mathbf{x}'))\|_2 \\ &\le \varepsilon_{\mathrm{al}} + L_f\delta + \varepsilon_{\mathrm{al}} = h_{\mathrm{aug}}. \end{aligned} \tag{65}$$

Since the identity belongs to $\mathcal{T}$, this includes the unaugmented $\mathbf{x}$. For fixed $\mathbf{x} \in C_k^\circ \cap \mathcal{V}$, average over an independent $\mathbf{X}' \sim P_k$ and $A' \sim \mathrm{Unif}(\mathcal{T})$. Every representation has unit norm, so distances involving an observation outside $C_k^\circ \cap \mathcal{V}$ are at most two. Hence, by (33) and Jensen's inequality,

$$\begin{aligned} \|f(\mathbf{x}) - \mathbf{m}_k\|_2 &= \|f(\mathbf{x}) - \mathbb{E}_{\mathbf{X}',A'}[f(A'(\mathbf{X}'))]\|_2 \\ &= \|\mathbb{E}_{\mathbf{X}',A'}[f(\mathbf{x}) - f(A'(\mathbf{X}'))]\|_2 \\ &\le \mathbb{E}_{\mathbf{X}',A'}\|f(\mathbf{x}) - f(A'(\mathbf{X}'))\|_2 \\ &\le P_k(C_k^\circ \cap \mathcal{V})h_{\mathrm{aug}} + 2P_k(\{C_k^\circ \cap \mathcal{V}\}^c) \\ &\le h_{\mathrm{aug}} + 2b_{\mathrm{aug}}. \end{aligned} \tag{66}$$

This proves (63). For (62), first keep the conditioning explicit. Because $P_k = \mathrm{Law}(\mathbf{X} \mid Y = k)$ and $\mathbf{Z}_0 = f(A(\mathbf{X}))$, take independent $\mathbf{X}, \mathbf{X}' \sim P_k$ and independent $A, A' \sim \mathrm{Unif}(\mathcal{T})$. Equation (33) equivalently gives $\mathbf{m}_k = \mathbb{E}_{\mathbf{X}',A'}[f(A'(\mathbf{X}'))]$. Therefore, by linearity of expectation and Jensen's inequality,

$$\begin{aligned} \mathbb{E}[\|\mathbf{Z}_0 - \mathbf{m}_k\|_2 \mid Y = k] &= \mathbb{E}[\|f(A(\mathbf{X})) - \mathbf{m}_k\|_2 \mid Y = k] \\ &= \mathbb{E}_{\mathbf{X}\sim P_k,A}\left[\|f(A(\mathbf{X})) - \mathbf{m}_k\|_2\right] \\ &= \mathbb{E}_{\mathbf{X}\sim P_k,A}\left[\|f(A(\mathbf{X})) - \mathbb{E}_{\mathbf{X}',A'}[f(A'(\mathbf{X}'))]\|_2\right] \\ &= \mathbb{E}_{\mathbf{X}\sim P_k,A}\left[\|\mathbb{E}_{\mathbf{X}',A'}[f(A(\mathbf{X})) - f(A'(\mathbf{X}'))]\|_2\right] \\ &\le \mathbb{E}_{\mathbf{X},\mathbf{X}',A,A'}\|f(A(\mathbf{X})) - f(A'(\mathbf{X}'))\|_2. \end{aligned}$$

If both underlying observations lie in $C_k^\circ \cap \mathcal{V}$, then (65) bounds the last distance by h_{aug}. Otherwise at least one observation lies outside this set, and the unit-norm constraint bounds the distance by two.

Splitting the expectation into these two cases, and then applying the union bound, gives

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X},\mathbf{X}',A,A'}\|f(A(\mathbf{X})) - f(A'(\mathbf{X}'))\|_2 \\
&\le h_{\mathrm{aug}} P(\mathbf{X} \in C_k^\circ \cap \mathcal{V},\ \mathbf{X}' \in C_k^\circ \cap \mathcal{V}) \\
&\quad + 2P(\ \mathbf{X} \notin C_k^\circ \cap \mathcal{V} \text{ or } \mathbf{X}' \notin C_k^\circ \cap \mathcal{V}\ ) \\
&\le h_{\mathrm{aug}} + 2P_k(\{C_k^\circ \cap \mathcal{V}\}^c) + 2P_k(\{C_k^\circ \cap \mathcal{V}\}^c) \\
&\le h_{\mathrm{aug}} + 4b_{\mathrm{aug}}. \qquad (67)
\end{aligned}$$

□

The endpoint estimate below uses the following standard nonhomogeneous integral form of Gronwall's inequality, which we state without proof.

Lemma 5 (Integral Gronwall inequality)**.** *Let $h : [0,1] \to [0,\infty)$ be continuous, and let $a, b : [0,1] \to [0,\infty)$ be integrable and $c \ge 0$. If*

$$h(t) \le c + \int_0^t a(u)h(u)\,\mathrm{d}u + \int_0^t b(u)\,\mathrm{d}u, \qquad t \in [0,1],$$

then

$$h(t) \le \exp\left(\int_0^t a(u)\,\mathrm{d}u\right)\left(c + \int_0^t b(s)\,\mathrm{d}s\right).$$

The same endpoint estimate also exchanges an expectation and a time integral through the following standard result, stated here without proof.

Lemma 6 (Tonelli's theorem for nonnegative integrands)**.** *Let $H(\omega, t)$ be a nonnegative jointly measurable function on a probability space times $[0,1]$. Then*

$$\mathbb{E}\left[\int_0^1 H(\omega,t)\,\mathrm{d}t\right] = \int_0^1 \mathbb{E}[H(\omega,t)]\,\mathrm{d}t.$$

Lemma 7 (Velocity regression controls terminal Wasserstein error)**.** *Under Assumption 2,*

$$W_1(\widehat{p}_1, \nu_\varepsilon) \le \left(\mathbb{E}_{\Pi_f}[\|\widehat{\mathbf{Z}}_1 - \mathbf{R}\|_2^2]\right)^{1/2} \le e^{L_v}\sqrt{\frac{d^\star}{q_{\min}}\mathcal{R}_{\mathrm{FM}}}, \qquad (68)$$

and

$$\|\Phi_1(\mathbf{z}) - \Phi_1(\mathbf{z}')\|_2 \le e^{L_v}\|\mathbf{z} - \mathbf{z}'\|_2. \qquad (69)$$

Proof. The learned trajectory and analytic path start from the same representation $\mathbf{Z}_0$. By (25) and (11), respectively, their derivatives are $\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\mathbf{Z}}_t = \boldsymbol{v}_\phi(\widehat{\mathbf{Z}}_t, t)$ and $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Z}_t = \mathbf{U}_t$. Because both paths are absolutely continuous, the fundamental theorem of calculus on $[0,t]$ gives their integral forms

$$\begin{aligned}
\widehat{\mathbf{Z}}_t &= \mathbf{Z}_0 + \int_0^t \boldsymbol{v}_\phi(\widehat{\mathbf{Z}}_u, u)\,\mathrm{d}u, \\
\mathbf{Z}_t &= \mathbf{Z}_0 + \int_0^t \mathbf{U}_u\,\mathrm{d}u.
\end{aligned}$$

Subtracting these identities and adding and subtracting $\boldsymbol{v}_\phi(\mathbf{Z}_u, u)$ inside the integrand gives

$$\begin{aligned}
\widehat{\mathbf{Z}}_t - \mathbf{Z}_t &= \int_0^t [\boldsymbol{v}_\phi(\widehat{\mathbf{Z}}_u, u) - \mathbf{U}_u]\,\mathrm{d}u \\
&= \int_0^t [\boldsymbol{v}_\phi(\widehat{\mathbf{Z}}_u, u) - \boldsymbol{v}_\phi(\mathbf{Z}_u, u)]\,\mathrm{d}u + \int_0^t [\boldsymbol{v}_\phi(\mathbf{Z}_u, u) - \mathbf{U}_u]\,\mathrm{d}u.
\end{aligned}$$

Taking norms, applying the triangle inequality for integrals, and then using the spatial Lipschitz condition in Assumption 2 gives

$$\|\widehat{\mathbf{Z}}_t - \mathbf{Z}_t\|_2 \le \int_0^t \ell_\phi(u)\|\widehat{\mathbf{Z}}_u - \mathbf{Z}_u\|_2\,\mathrm{d}u + \int_0^t \|\boldsymbol{v}_\phi(\mathbf{Z}_u, u) - \mathbf{U}_u\|_2\,\mathrm{d}u. \qquad (70)$$

Applying Lemma 5 with $h(t) = \|\widehat{\mathbf{Z}}_t - \mathbf{Z}_t\|_2$, $a(t) = \ell_\phi(t)$, and $b(t) = \|\boldsymbol{v}_\phi(\mathbf{Z}_t, t) - \mathbf{U}_t\|_2$, $c = 0$, and using $\int_0^1 \ell_\phi(u)\,\mathrm{d}u \le L_v$ from Assumption 2, yields

$$\|\widehat{\mathbf{Z}}_t - \mathbf{Z}_t\|_2 \le \exp\left(\int_0^t \ell_\phi(u)\,\mathrm{d}u\right) \int_0^t \|\boldsymbol{v}_\phi(\mathbf{Z}_u, u) - \mathbf{U}_u\|_2\,\mathrm{d}u.$$

At $t = 1$, (10) gives $\mathbf{Z}_1 = \mathbf{R}$; hence

$$\|\widehat{\mathbf{Z}}_1 - \mathbf{R}\|_2 \le e^{L_v} \int_0^1 \|\boldsymbol{v}_\phi(\mathbf{Z}_t, t) - \mathbf{U}_t\|_2\,\mathrm{d}t. \tag{71}$$

Squaring (71), applying Cauchy–Schwarz on $[0, 1]$, and then applying Lemma 6 gives

$$\begin{aligned}
&\mathbb{E}_{\Pi_f}[\|\widehat{\mathbf{Z}}_1 - \mathbf{R}\|_2^2] \\
&\le e^{2L_v}\mathbb{E}_{\Pi_f}\left[\left(\int_0^1 \|\boldsymbol{v}_\phi(\mathbf{Z}_t, t) - \mathbf{U}_t\|_2\,\mathrm{d}t\right)^2\right] \\
&\le e^{2L_v}\mathbb{E}_{\Pi_f}\left[\left(\int_0^1 1^2\,\mathrm{d}t\right)\left(\int_0^1 \|\boldsymbol{v}_\phi(\mathbf{Z}_t, t) - \mathbf{U}_t\|_2^2\,\mathrm{d}t\right)\right] \\
&= e^{2L_v}\int_0^1 \mathbb{E}_{\Pi_f}[\|\boldsymbol{v}_\phi(\mathbf{Z}_t, t) - \mathbf{U}_t\|_2^2]\,\mathrm{d}t.
\end{aligned}$$

On the other hand, the definition (20) and the density q_T imply

$$\begin{aligned}
d^\star\mathcal{R}_{\mathrm{FM}} &= \mathbb{E}_{\Pi_f,T}[\|\boldsymbol{v}_\phi(\mathbf{Z}_T, T) - \mathbf{U}_T\|_2^2] \\
&= \int_0^1 q_T(t)\mathbb{E}_{\Pi_f}[\|\boldsymbol{v}_\phi(\mathbf{Z}_t, t) - \mathbf{U}_t\|_2^2]\,\mathrm{d}t \\
&\ge q_{\min}\int_0^1 \mathbb{E}_{\Pi_f}[\|\boldsymbol{v}_\phi(\mathbf{Z}_t, t) - \mathbf{U}_t\|_2^2]\,\mathrm{d}t,
\end{aligned}$$

where the last inequality uses the time-coverage condition $q_T(t) \ge q_{\min}$ in Assumption 2. Combining the two displays yields

$$\mathbb{E}_{\Pi_f}[\|\widehat{\mathbf{Z}}_1 - \mathbf{R}\|_2^2] \le \frac{d^\star e^{2L_v}}{q_{\min}}\mathcal{R}_{\mathrm{FM}}. \tag{72}$$

By (25) and the endpoint marginal of Π_f, the pair $(\widehat{\mathbf{Z}}_1, \mathbf{R})$ is a coupling of $\widehat{p}_1$ and ν_ε. Using this admissible coupling in the infimum defining W_1 in (26), followed by Cauchy–Schwarz and the preceding endpoint second-moment bound, gives

$$\begin{aligned}
W_1(\widehat{p}_1, \nu_\varepsilon) &\le \mathbb{E}_{\Pi_f}[\|\widehat{\mathbf{Z}}_1 - \mathbf{R}\|_2] \\
&\le \left(\mathbb{E}_{\Pi_f}[\|\widehat{\mathbf{Z}}_1 - \mathbf{R}\|_2^2]\right)^{1/2} \\
&\le e^{L_v}\sqrt{\frac{d^\star}{q_{\min}}\mathcal{R}_{\mathrm{FM}}}.
\end{aligned}$$

It remains to prove (69). Integrating the two instances of the learned ODE (25), subtracting them, and using the spatial Lipschitz condition in Assumption 2 yields

$$\|\Phi_t(\mathbf{z}) - \Phi_t(\mathbf{z}')\|_2 \le \|\mathbf{z} - \mathbf{z}'\|_2 + \int_0^t \ell_\phi(u)\|\Phi_u(\mathbf{z}) - \Phi_u(\mathbf{z}')\|_2\,\mathrm{d}u.$$

Applying Lemma 5 with $c = \|\mathbf{z} - \mathbf{z}'\|_2$ and $b = 0$ gives

$$\|\Phi_t(\mathbf{z}) - \Phi_t(\mathbf{z}')\|_2 \le \exp\left(\int_0^t \ell_\phi(u)\,\mathrm{d}u\right)\|\mathbf{z} - \mathbf{z}'\|_2 \le e^{L_v}\|\mathbf{z} - \mathbf{z}'\|_2.$$

Taking $t = 1$ proves (69). □

Lemma 7 controls the mismatch between the overall terminal law and the reference law, but downstream classification requires a class-wise conclusion. Under transport identifiability, the next lemma makes this conversion: each terminal class mean is close to its assigned reference component mean, so their pairwise geometry inherits that of the reference components up to an error controlled by the terminal Wasserstein distance. Here, $\Gamma^\star$ is the optimal terminal–reference coupling introduced after (26), and $\tau^\star$ is the optimal class–component permutation defined in (29).

Lemma 8 (Leakage controls terminal class-centre geometry)**.** *Define the terminal and reference component means by*

$$\widehat{\mathbf{m}}_k := \mathbb{E}[\widehat{\mathbf{Z}}_1 \mid Y = k], \qquad \mathbf{b}_j := \int \mathbf{r}\, Q_j^\varepsilon(\mathrm{d}\mathbf{r}), \qquad D := \frac{1 + 2C_{\mathrm{id}}}{p_{\min}} W_1(\widehat{p}_1, \nu_\varepsilon). \tag{73}$$

Under Assumption 3,

$$\max_k \|\widehat{\mathbf{m}}_k - \mathbf{b}_{\tau^\star(k)}\|_2 \le D, \tag{74}$$

and

$$\max_{k \neq \ell} \big| \|\widehat{\mathbf{m}}_k - \widehat{\mathbf{m}}_\ell\|_2 - \|\mathbf{b}_{\tau^\star(k)} - \mathbf{b}_{\tau^\star(\ell)}\|_2 \big| \le 2D. \tag{75}$$

Remark 1 (Reference center versus component mean)**.** *Since Q_j^ε is rotationally symmetric around $\mathbf{c}_j$, its mean satisfies $\mathbf{b}_j = \alpha_\varepsilon \mathbf{c}_j$ for some $\alpha_\varepsilon \in [0,1]$ that is common to all j. Thus $\mathbf{b}_j$ has the same direction as $\mathbf{c}_j$ but generally lies inside the unit sphere. For example, on the unit circle with $\mathbf{c}_j = (1,0)$, the symmetric points $(\cos\theta, \sin\theta)$ and $(\cos\theta, -\sin\theta)$ average to $(\cos\theta, 0) = \cos\theta\, \mathbf{c}_j$. In particular, $\alpha_\varepsilon = 1$ when $\varepsilon = 0$.*

Proof of Lemma 8. We adapt the component-mean decomposition in Jiao et al. (2026). Recall that $p_k = P(Y = k)$, w_j is the mixture weight of Q_j^ε in ν_ε, and $\pi_k(\widehat{\mathbf{z}}) = P(Y = k \mid \widehat{\mathbf{Z}}_1 = \widehat{\mathbf{z}})$. Set $F_k := \mathcal{S}^\varepsilon_{\tau^\star(k)}$, the support of the reference component assigned to class k. In the integrals below, $\mathrm{d}\Gamma^\star$ abbreviates $\Gamma^\star(\mathrm{d}\widehat{\mathbf{z}}, \mathrm{d}\mathbf{r})$.

First, the $\widehat{\mathbf{Z}}_1$-marginal of $\Gamma^\star$ is $\widehat{p}_1$. Bayes' rule in Radon–Nikodym form gives

$$\frac{\mathrm{d}P(\widehat{\mathbf{Z}}_1 \in \cdot \mid Y = k)}{\mathrm{d}\widehat{p}_1}(\widehat{\mathbf{z}}) = \frac{\pi_k(\widehat{\mathbf{z}})}{p_k}.$$

Starting from the definition of $\widehat{\mathbf{m}}_k$ and using this derivative, followed by the $\widehat{\mathbf{Z}}_1$-marginal identity, yields

$$\begin{aligned}
\widehat{\mathbf{m}}_k &= \int \widehat{\mathbf{z}}\, P(\widehat{\mathbf{Z}}_1 \in \mathrm{d}\widehat{\mathbf{z}} \mid Y = k) \\
&= \frac{1}{p_k} \int \pi_k(\widehat{\mathbf{z}}) \widehat{\mathbf{z}}\, \widehat{p}_1(\mathrm{d}\widehat{\mathbf{z}}) \\
&= \frac{1}{p_k} \int \pi_k(\widehat{\mathbf{z}}) \widehat{\mathbf{z}}\, \mathrm{d}\Gamma^\star.
\end{aligned}$$

Equivalently, $p_k \widehat{\mathbf{m}}_k = \int \pi_k(\widehat{\mathbf{z}}) \widehat{\mathbf{z}}\, \mathrm{d}\Gamma^\star$. Second, the $\mathbf{R}^\star$-marginal of $\Gamma^\star$ is $\nu_\varepsilon = \sum_j w_j Q_j^\varepsilon$. Since the component supports are disjoint, the relevant Radon–Nikodym derivative is

$$\frac{\mathrm{d}Q^\varepsilon_{\tau^\star(k)}}{\mathrm{d}\nu_\varepsilon}(\mathbf{r}) = \frac{\mathbf{1}\{\mathbf{r} \in F_k\}}{w_{\tau^\star(k)}} \qquad \nu_\varepsilon\text{-a.e.}$$

Starting from the definition of $\mathbf{b}_{\tau^\star(k)}$, using this derivative and then the $\mathbf{R}^\star$-marginal identity gives

$$\begin{aligned}
\mathbf{b}_{\tau^\star(k)} &= \int \mathbf{r}\, Q^\varepsilon_{\tau^\star(k)}(\mathrm{d}\mathbf{r}) \\
&= \frac{1}{w_{\tau^\star(k)}} \int_{F_k} \mathbf{r}\, \nu_\varepsilon(\mathrm{d}\mathbf{r}) \\
&= \frac{1}{w_{\tau^\star(k)}} \int \mathbf{1}\{\mathbf{r} \in F_k\} \mathbf{r}\, \mathrm{d}\Gamma^\star.
\end{aligned}$$

Equivalently, $w_{\tau^\star(k)}\mathbf{b}_{\tau^\star(k)} = \int \mathbf{1}\{\mathbf{r} \in F_k\}\mathbf{r}\,\mathrm{d}\Gamma^\star$. Subtracting these two identities, and then adding and subtracting $\pi_k(\widehat{\mathbf{z}})\mathbf{r}$ inside the integrand, yields

$$\begin{aligned} p_k\widehat{\mathbf{m}}_k - w_{\tau^\star(k)}\mathbf{b}_{\tau^\star(k)} &= \int [\pi_k(\widehat{\mathbf{z}})\widehat{\mathbf{z}} - \mathbf{1}\{\mathbf{r} \in F_k\}\mathbf{r}]\ \mathrm{d}\Gamma^\star \\ &= \int \pi_k(\widehat{\mathbf{z}})(\widehat{\mathbf{z}} - \mathbf{r})\,\mathrm{d}\Gamma^\star \\ &\quad + \int [\pi_k(\widehat{\mathbf{z}}) - \mathbf{1}\{\mathbf{r} \in F_k\}]\,\mathbf{r}\,\mathrm{d}\Gamma^\star. \end{aligned} \tag{76}$$

Write $\ell_k := p_k + w_{\tau^\star(k)} - 2q_k^\star$. Since $\pi_k(\widehat{\mathbf{z}}) \in [0,1]$ and $\mathbf{1}\{\mathbf{r} \in F_k\} \in \{0,1\}$, pointwise we have

$$|\pi_k(\widehat{\mathbf{z}}) - \mathbf{1}\{\mathbf{r} \in F_k\}| = \pi_k(\widehat{\mathbf{z}}) + \mathbf{1}\{\mathbf{r} \in F_k\} - 2\pi_k(\widehat{\mathbf{z}})\mathbf{1}\{\mathbf{r} \in F_k\}.$$

Recall from the two marginals of $\Gamma^\star$ and (28) that

$$\int \pi_k(\widehat{\mathbf{z}})\,\mathrm{d}\Gamma^\star = p_k,$$

$$\int \mathbf{1}\{\mathbf{r} \in F_k\}\,\mathrm{d}\Gamma^\star = w_{\tau^\star(k)},$$

$$\int \pi_k(\widehat{\mathbf{z}})\mathbf{1}\{\mathbf{r} \in F_k\}\,\mathrm{d}\Gamma^\star = q_k^\star.$$

Therefore, integrating the pointwise identity gives

$$\int |\pi_k(\widehat{\mathbf{z}}) - \mathbf{1}\{\mathbf{r} \in F_k\}|\ \mathrm{d}\Gamma^\star = p_k + w_{\tau^\star(k)} - 2q_k^\star = \ell_k.$$

Consequently, the second term in (76) satisfies

$$\left\| \int [\pi_k(\widehat{\mathbf{z}}) - \mathbf{1}\{\mathbf{r} \in F_k\}]\,\mathbf{r}\,\mathrm{d}\Gamma^\star \right\|_2 \le \int |\pi_k(\widehat{\mathbf{z}}) - \mathbf{1}\{\mathbf{r} \in F_k\}|\,\|\mathbf{r}\|_2\,\mathrm{d}\Gamma^\star = \ell_k,$$

where the last equality uses $\|\mathbf{r}\|_2 = 1$. Similarly, the first term satisfies

$$\begin{aligned} \left\| \int \pi_k(\widehat{\mathbf{z}})(\widehat{\mathbf{z}} - \mathbf{r})\,\mathrm{d}\Gamma^\star \right\|_2 &\le \int \pi_k(\widehat{\mathbf{z}})\|\widehat{\mathbf{z}} - \mathbf{r}\|_2\,\mathrm{d}\Gamma^\star \\ &\le \int \|\widehat{\mathbf{z}} - \mathbf{r}\|_2\,\mathrm{d}\Gamma^\star = W_1(\widehat{p}_1, \nu_\varepsilon), \end{aligned}$$

because $\Gamma^\star$ is an optimal coupling. Combining the two bounds yields

$$\|p_k\widehat{\mathbf{m}}_k - w_{\tau^\star(k)}\mathbf{b}_{\tau^\star(k)}\|_2 \le W_1(\widehat{p}_1, \nu_\varepsilon) + \ell_k. \tag{77}$$

Moreover, $|p_k - w_{\tau^\star(k)}| \le \ell_k$ and $\|\mathbf{b}_{\tau^\star(k)}\|_2 \le 1$. Consequently,

$$\begin{aligned} p_k\|\widehat{\mathbf{m}}_k - \mathbf{b}_{\tau^\star(k)}\|_2 &\le W_1(\widehat{p}_1, \nu_\varepsilon) + 2\ell_k \\ &\le (1 + 2C_{\mathrm{id}})W_1(\widehat{p}_1, \nu_\varepsilon). \end{aligned} \tag{78}$$

Division by $p_k \ge p_{\min}$ proves (74). Finally, for $k \ne \ell$, the reverse triangle inequality followed by (74) gives

$$\begin{aligned} \big|\|\widehat{\mathbf{m}}_k - \widehat{\mathbf{m}}_\ell\|_2 - \|\mathbf{b}_{\tau^\star(k)} - \mathbf{b}_{\tau^\star(\ell)}\|_2\big| &\le \|(\widehat{\mathbf{m}}_k - \mathbf{b}_{\tau^\star(k)}) - (\widehat{\mathbf{m}}_\ell - \mathbf{b}_{\tau^\star(\ell)})\|_2 \\ &\le \|\widehat{\mathbf{m}}_k - \mathbf{b}_{\tau^\star(k)}\|_2 + \|\widehat{\mathbf{m}}_\ell - \mathbf{b}_{\tau^\star(\ell)}\|_2 \\ &\le 2D. \end{aligned}$$

This proves (75) and completes the proof. □

Since $\|\mathbf{b}_j - \mathbf{c}_j\|_2 \le r_\varepsilon$, Lemma 8 gives the terminal separation bound

$$\min_{k \ne \ell} \|\widehat{\mathbf{m}}_k - \widehat{\mathbf{m}}_\ell\|_2 \ge \Delta_{\mathcal{C}} - 2r_\varepsilon - 2D. \tag{79}$$

By Lemma 7,

$$D \le \frac{(1 + 2C_{\mathrm{id}})e^{L_v}}{p_{\min}} \sqrt{\frac{d^\star}{q_{\min}}\mathcal{R}_{\mathrm{FM}}}. \tag{80}$$

Theorem 1 (Population FBDM loss controls classification error)**.** *Suppose Assumptions 1–3 hold. Fix a view-stability tolerance $\varepsilon_{\mathrm{al}} > 0$ such that*

$$e^{-L_v}(\Delta_{\mathcal{C}} - 2r_\varepsilon) > 4L_f\delta + 8\varepsilon_{\mathrm{al}} + 12(1-\sigma).$$

Define

$$\Delta_0 := e^{-L_v}(\Delta_{\mathcal{C}} - 2r_\varepsilon), \qquad C_{\mathrm{err}} := \frac{N_{\mathrm{aug}}}{\varepsilon_{\mathrm{al}}}\sqrt{\frac{d^\star}{\lambda}}, \tag{81}$$

$$C_{\mathrm{sep}} := \frac{2(1+2C_{\mathrm{id}})}{p_{\min}}\sqrt{\frac{d^\star}{q_{\min}}} + \frac{12C_{\mathrm{err}}}{p_{\min}}, \tag{82}$$

$$g_{\mathrm{aug}} := \Delta_0 - 4L_f\delta - 8\varepsilon_{\mathrm{al}} - 12(1-\sigma). \tag{83}$$

Then $g_{\mathrm{aug}} > 0$. Set

$$r_0 := \left(\frac{g_{\mathrm{aug}}}{2C_{\mathrm{sep}}}\right)^2. \tag{84}$$

These constants depend only on the stated uniform bounds. If $\mathcal{R}_{\mathrm{FBDM}}(f,\phi) \le r_0$, then

$$\mathrm{Err}(G_f) \le (1-\sigma) + C_{\mathrm{err}}\sqrt{\mathcal{R}_{\mathrm{FBDM}}(f,\phi)}.$$

Proof of Theorem 1. By construction and (33), $\mathrm{Law}(\mathbf{Z}_0 \mid Y=k) = f_\# P_{\mathcal{A},k}$ and $\mathbf{m}_k = \mathbb{E}[\mathbf{Z}_0 \mid Y=k]$, where $\mathbf{Z}_0$ is the source representation in the coupling Π_f. Let Φ_1 denote the terminal map of the learned flow and define the corresponding terminal class mean by $\widehat{\mathbf{m}}_k := \mathbb{E}[\Phi_1(\mathbf{Z}_0) \mid Y=k]$. For readability, recall the three quantities that measure within-class spread:

$$\mathcal{V} := \left\{\mathbf{x} \in \mathrm{supp}(P_X) : \max_{A,A' \in \mathcal{T}} \|f(A(\mathbf{x})) - f(A'(\mathbf{x}))\|_2 \le \varepsilon_{\mathrm{al}}\right\},$$
$$\mathcal{U} := P_X(\mathcal{V}^c),$$
$$h_{\mathrm{aug}} := L_f\delta + 2\varepsilon_{\mathrm{al}},$$
$$b_{\mathrm{aug}} := 1 - \sigma + \frac{\mathcal{U}}{p_{\min}}. \tag{85}$$

Here $\mathcal{U}$ is the mass of observations whose augmented views are not $\varepsilon_{\mathrm{al}}$-stable, h_{aug} bounds the diameter of a stable class core, and b_{aug} bounds the mass not covered by that core. These are exactly the quantities entering the class-spread bounds in Lemma 4. The next bound compares evolving the class mean, $\Phi_1(\mathbf{m}_k)$, with averaging the individually evolved representations, $\widehat{\mathbf{m}}_k$. Since $\Phi_1(\mathbf{m}_k)$ is deterministic conditional on $Y=k$, conditional Jensen's inequality gives

$$\begin{aligned}
\|\Phi_1(\mathbf{m}_k) - \widehat{\mathbf{m}}_k\|_2 &= \|\mathbb{E}[\Phi_1(\mathbf{m}_k) - \Phi_1(\mathbf{Z}_0) \mid Y=k]\|_2 \\
&\le \mathbb{E}[\|\Phi_1(\mathbf{m}_k) - \Phi_1(\mathbf{Z}_0)\|_2 \mid Y=k] \\
&\le e^{L_v}\mathbb{E}[\|\mathbf{m}_k - \mathbf{Z}_0\|_2 \mid Y=k] \\
&\le e^{L_v}(h_{\mathrm{aug}} + 4b_{\mathrm{aug}}).
\end{aligned} \tag{86}$$

Here the penultimate inequality uses the flow stability in Lemma 7, and the last uses the class-spread bound from Lemma 4. Although $\mathbf{m}_k$ may lie inside the unit sphere, $\Phi_1(\mathbf{m}_k)$ is well-defined because the flow acts on $\mathbb{R}^{d^\star}$; (86) controls its discrepancy from $\widehat{\mathbf{m}}_k$.

On the reference side, define

$$\mathbf{b}_j := \int \mathbf{r}\, Q_j^\varepsilon(\mathrm{d}\mathbf{r}), \qquad D := \frac{1+2C_{\mathrm{id}}}{p_{\min}} W_1(\widehat{p}_1, \nu_\varepsilon). \tag{87}$$

Here $\mathbf{b}_j$ is the mean of reference component j, and Lemma 8 shows that D uniformly bounds $\|\widehat{\mathbf{m}}_k - \mathbf{b}_{\tau^\star(k)}\|_2$, where $\tau^\star(k)$ is the reference component matched to class k. Also recall the reference-centre gap and component radius

$$\Delta_{\mathcal{C}} := \min_{i \ne j} \|\mathbf{c}_i - \mathbf{c}_j\|_2, \qquad r_\varepsilon := \max_j \sup_{\mathbf{r} \in \mathcal{S}_j^\varepsilon} \|\mathbf{r} - \mathbf{c}_j\|_2. \tag{88}$$

The radius definition implies $\|\mathbf{b}_j - \mathbf{c}_j\|_2 \le r_\varepsilon$. For $k \neq \ell$, Lemma 8 thus gives

$$\begin{aligned}\Delta_{\mathcal{C}} - 2r_\varepsilon &\le \|\mathbf{b}_{\tau^\star(k)} - \mathbf{b}_{\tau^\star(\ell)}\|_2 \\ &\le 2D + \|\widehat{\mathbf{m}}_k - \widehat{\mathbf{m}}_\ell\|_2 \\ &\le 2D + e^{L_v}\{2h_{\text{aug}} + 8b_{\text{aug}} + \|\mathbf{m}_k - \mathbf{m}_\ell\|_2\}. \end{aligned} \tag{89}$$

Define the initial class-mean gap and the reference gap that remains after component perturbation and inverse-flow contraction by

$$\Delta_f := \min_{k\neq\ell} \|\mathbf{m}_k - \mathbf{m}_\ell\|_2, \qquad \Delta_0 := e^{-L_v}(\Delta_{\mathcal{C}} - 2r_\varepsilon). \tag{90}$$

The preceding display then yields

$$\Delta_f \ge \Delta_0 - 2e^{-L_v}D - 2h_{\text{aug}} - 8b_{\text{aug}}. \tag{91}$$

To compare this separation with the stable-core radius, define

$$\begin{aligned} g_{\text{aug}} &:= \Delta_0 - 4L_f\delta - 8\varepsilon_{\text{al}} - 12(1-\sigma), \\ C_{\text{sep}} &:= \frac{2(1+2C_{\text{id}})}{p_{\min}}\sqrt{\frac{d^\star}{q_{\min}}} + \frac{12C_{\text{err}}}{p_{\min}}. \end{aligned} \tag{92}$$

Thus g_{aug} is the fixed separation margin left after the augmentation terms, whereas C_{sep} collects the coefficients of the loss-dependent errors. Here $p_{\min} = \min_k P(Y=k)$, $q_{\min}$ is the lower bound on the flow-time density, and C_{id} is the identifiability constant in Assumption 3; C_{err} is given in Theorem 1 above. To prove that an image is closer to its own class mean, fix $\mathbf{x} \in C_k^\circ \cap \mathcal{V}$ and $\ell \neq k$. We show that the difference between its distances to the other class mean and its own class mean is positive. The triangle inequality, the within-class radius bound (63), and the class-mean separation bound (91) give the first three inequalities below. We then substitute the bounds (56) and (80):

$$\begin{aligned}\|f(\mathbf{x}) - \mathbf{m}_\ell\|_2 - \|f(\mathbf{x}) - \mathbf{m}_k\|_2 &\ge \Delta_f - 2\|f(\mathbf{x}) - \mathbf{m}_k\|_2 \\ &\ge \Delta_f - 2(h_{\text{aug}} + 2b_{\text{aug}}) \\ &\ge \Delta_0 - 2e^{-L_v}D - 4h_{\text{aug}} - 12b_{\text{aug}} \\ &= \Delta_0 - 2e^{-L_v}D - 4(L_f\delta + 2\varepsilon_{\text{al}}) - 12\left(1 - \sigma + \frac{\mathcal{U}}{p_{\min}}\right) \\ &= \{\Delta_0 - 4L_f\delta - 8\varepsilon_{\text{al}} - 12(1-\sigma)\} - 2e^{-L_v}D - \frac{12\mathcal{U}}{p_{\min}} \\ &= g_{\text{aug}} - 2e^{-L_v}D - \frac{12\mathcal{U}}{p_{\min}} \\ &\ge g_{\text{aug}} - \frac{2(1+2C_{\text{id}})}{p_{\min}}\sqrt{\frac{d^\star}{q_{\min}}\mathcal{R}_{\text{FM}}} - \frac{12\mathcal{U}}{p_{\min}} \\ &\ge g_{\text{aug}} - \frac{2(1+2C_{\text{id}})}{p_{\min}}\sqrt{\frac{d^\star}{q_{\min}}\mathcal{R}_{\text{FM}}} - \frac{12C_{\text{err}}}{p_{\min}}\sqrt{\mathcal{R}_{\text{FBDM}}} \\ &\ge g_{\text{aug}} - C_{\text{sep}}\sqrt{\mathcal{R}_{\text{FBDM}}} \ge \frac{g_{\text{aug}}}{2} > 0. \end{aligned} \tag{93}$$

The first equality substitutes the definitions of h_{aug} and b_{aug} from (85); the next equality identifies the bracketed fixed margin as g_{aug}. The following two inequalities respectively use the bounds on D in (80) and on $\mathcal{U}$ in (56). Finally, $\mathcal{R}_{\text{FM}} \le \mathcal{R}_{\text{FBDM}}$ and the definition of C_{sep} collect both loss-dependent terms, while (84) gives the last inequality. Therefore, for every $\ell \neq k$,

$$\|f(\mathbf{x}) - \mathbf{m}_\ell\|_2 > \|f(\mathbf{x}) - \mathbf{m}_k\|_2. \tag{94}$$

Hence $G_f(\mathbf{x}) = k$. Misclassification is confined to observations outside their class core or outside the stable-view set:

$$\begin{aligned}\text{Err}(G_f) &\le \sum_{k=1}^{K} p_k P_k((C_k^\circ)^c) + \mathcal{U} \\ &\le (1-\sigma) + \mathcal{U} \le (1-\sigma) + C_{\text{err}}\sqrt{\mathcal{R}_{\text{FBDM}}}. \end{aligned} \tag{95}$$

This proves (35). Augmentation quality contributes the uncovered mass $1-\sigma$ and the margin g_{aug}; neither leakage nor Wasserstein discrepancy remains as an uncontrolled term in the final loss-to-error bound. □

C.4 A Geometric Sufficient Condition for Leakage Control

Theorem 1 uses Assumption 3 to turn small terminal W_1 discrepancy into small class–component leakage. The purpose of this section is to give a concrete geometric condition under which that implication holds. The condition allows a controlled fraction of terminal mass to lie in or near an off-matched reference component.

Recall that $\mathcal{S}_j^\varepsilon := \mathrm{supp}(Q_j^\varepsilon)$ is the support of reference component j. For a distance threshold $\Delta > 0$, define

$$q_{kj}^{\star,<\Delta} := \int \pi_k(\widehat{\mathbf{z}})\mathbf{1}\{\mathbf{r} \in \mathcal{S}_j^\varepsilon\}\mathbf{1}\{\|\widehat{\mathbf{z}} - \mathbf{r}\|_2 < \Delta\}\,\Gamma^\star(\mathrm{d}\widehat{\mathbf{z}}, \mathrm{d}\mathbf{r}). \tag{96}$$

Thus $q_{kj}^{\star,<\Delta}$ is the part of $q_{kj}^\star$ whose terminal representation is within Δ of its coupled target in component j. When $j \neq \tau^\star(k)$, it measures class-k mass that is paired with an off-matched component and is already inside or close to that component.

Proposition 9 (A soft off-component margin controls leakage)**.** *Suppose that, for some* $\Delta_{\mathrm{sep}} > 0$ *and* $\alpha_{\mathrm{sep}} \in [0, 1)$,

$$\sum_{k=1}^K \sum_{j \neq \tau^\star(k)} q_{kj}^{\star,<\Delta_{\mathrm{sep}}} \leq \alpha_{\mathrm{sep}} \sum_{k=1}^K \sum_{j \neq \tau^\star(k)} q_{kj}^\star. \tag{97}$$

Then

$$\mathrm{Leak}(\Gamma^\star) \leq \frac{2}{(1 - \alpha_{\mathrm{sep}})\Delta_{\mathrm{sep}}} W_1(\widehat{p}_1, \nu_\varepsilon). \tag{98}$$

Condition (97) allows some off-matched mass to enter or approach a wrong component, but limits its proportion to α_{sep}. Hence at least a fraction $1 - \alpha_{\mathrm{sep}}$ of the off-matched mass pays transport cost no less than Δ_{sep}, allowing W_1 to control the leakage.

Proof. We first bound the total off-matched mass

$$\sum_{k=1}^K \sum_{j \neq \tau^\star(k)} q_{kj}^\star \tag{99}$$

using W_1. Because $\Gamma^\star$ is an optimal coupling,

$$\begin{aligned}
W_1(\widehat{p}_1, \nu_\varepsilon) &= \int \|\widehat{\mathbf{z}} - \mathbf{r}\|_2 \,\Gamma^\star(\mathrm{d}\widehat{\mathbf{z}}, \mathrm{d}\mathbf{r}) \\
&\geq \sum_{k=1}^K \sum_{j \neq \tau^\star(k)} \int \pi_k(\widehat{\mathbf{z}})\mathbf{1}\{\mathbf{r} \in \mathcal{S}_j^\varepsilon\}\mathbf{1}\{\|\widehat{\mathbf{z}} - \mathbf{r}\|_2 \geq \Delta_{\mathrm{sep}}\}\|\widehat{\mathbf{z}} - \mathbf{r}\|_2 \,\Gamma^\star(\mathrm{d}\widehat{\mathbf{z}}, \mathrm{d}\mathbf{r}) \\
&\geq \Delta_{\mathrm{sep}} \sum_{k=1}^K \sum_{j \neq \tau^\star(k)} \left(q_{kj}^\star - q_{kj}^{\star,<\Delta_{\mathrm{sep}}}\right) \\
&\geq (1 - \alpha_{\mathrm{sep}})\Delta_{\mathrm{sep}} \sum_{k=1}^K \sum_{j \neq \tau^\star(k)} q_{kj}^\star.
\end{aligned} \tag{100}$$

Here the first inequality restricts a nonnegative integral to off-matched pairs at distance at least Δ_{sep}; the second uses the definition (96); and the last uses (97).

To see why (99) controls leakage, view $Q^\star = (q_{kj}^\star)_{k,j}$ as a class–component mass matrix. After matching class k with column $\tau^\star(k)$, the entries $q_{k,\tau^\star(k)}^\star$ form its matched diagonal, and (99) sums all off-diagonal entries. Since the row and column sums are $p_k = \sum_j q_{kj}^\star$ and $w_j = \sum_i q_{ij}^\star$, class-$k$ leakage is one off-diagonal row sum plus the corresponding off-diagonal column sum. Each is bounded by the sum of all off-diagonal entries; hence

$$\begin{aligned}
p_k + w_{\tau^\star(k)} - 2q_{k,\tau^\star(k)}^\star &= \sum_{j \neq \tau^\star(k)} q_{kj}^\star + \sum_{i \neq k} q_{i,\tau^\star(k)}^\star \\
&\leq 2 \sum_{i=1}^K \sum_{j \neq \tau^\star(i)} q_{ij}^\star.
\end{aligned} \tag{101}$$

Taking the maximum over k in (101) and using (100) gives

$$\mathrm{Leak}(\Gamma^\star) \le \frac{2}{(1-\alpha_{\mathrm{sep}})\Delta_{\mathrm{sep}}} W_1(\widehat{p}_1, \nu_\varepsilon). \tag{102}$$

This proves the claim. □

Thus Assumption 3 holds with $C_{\mathrm{id}} = 2/[(1-\alpha_{\mathrm{sep}})\Delta_{\mathrm{sep}}]$.

C.5 Relation to the Batchwise Population Coupling

Theorem 1 is stated for the abstract coupling Π_f, whose endpoint marginal is exactly ν_ε. In the implemented algorithm, however, endpoints are produced by minibatch assignment, and their marginal need not equal ν_ε because the resulting component occupancies may differ from the prescribed mixture weights. This section quantifies this endpoint-marginal discrepancy and propagates its effect through the classification guarantee in Theorem 1.

To make the population averaging explicit, first draw a training set $\mathcal{D}_N = (\mathbf{X}_i)_{i=1}^N \sim P_X^{\otimes N}$. Conditional on $\mathcal{D}_N$, draw a minibatch $\mathcal{B}$ according to the training sampler, generate its two views, and form their shared targets. The conditional expectation below also includes the augmentation and target-construction randomness. Let $N_k(\mathcal{B})$ be the number of images in $\mathcal{B}$ assigned to component k, and define

$$\begin{aligned} \bar{w}_{B,k}(\mathcal{D}_N) &:= \mathbb{E}_{\mathcal{B}|\mathcal{D}_N}\left[\frac{N_k(\mathcal{B})}{B}\right], \\ \bar{w}_{B,k} &:= \mathbb{E}_{\mathcal{D}_N}[\bar{w}_{B,k}(\mathcal{D}_N)], \\ \delta_{\mathrm{mix}} &:= \frac{1}{2}\sum_{k=1}^{K'} |\bar{w}_{B,k} - w_k|. \end{aligned} \tag{103}$$

Once the sampler, augmentation, target construction, encoder, and assignment rule are fixed, $\bar{w}_{B,k}$ is induced by the population distribution P_X; it is a defined quantity, not an additional assumption. After averaging over both $\mathcal{D}_N$ and $\mathcal{B} \mid \mathcal{D}_N$, let Π_B denote the resulting population representation–endpoint coupling and let $\nu_B := \mathrm{Law}(\mathbf{R})$ be its endpoint marginal. Because the endpoint perturbation conditional on component k follows Q_k^ε, $\nu_B = \sum_k \bar{w}_{B,k} Q_k^\varepsilon$, whereas $\nu_\varepsilon = \sum_k w_k Q_k^\varepsilon$. Coupling the common mass within each component and transporting only the unmatched mass over the unit sphere, whose diameter is two, gives

$$\begin{aligned} W_1(\nu_B, \nu_\varepsilon) &:= \inf_{\Gamma \in \mathcal{C}(\nu_B, \nu_\varepsilon)} \int \|\mathbf{r} - \mathbf{r}'\|_2 \, \Gamma(\mathrm{d}\mathbf{r}, \mathrm{d}\mathbf{r}') \\ &\le 2\delta_{\mathrm{mix}}, \end{aligned} \tag{104}$$

where $\mathcal{C}(\nu_B, \nu_\varepsilon)$ denotes the set of couplings of ν_B and ν_ε. Define

$$\mathcal{R}_{\mathrm{FM},B} := \frac{1}{d^\star}\mathbb{E}_{\Pi_B,T}\left[\|\boldsymbol{v}_\phi(\mathbf{Z}_T, T) - \mathbf{U}_T\|_2^2\right]. \tag{105}$$

The proof of Lemma 7 applies to this batchwise pairing. Writing $\widehat{p}_{1,B}$ for the law of $\Phi_1(\mathbf{Z}_0)$ under Π_B, the triangle inequality gives

$$W_1(\widehat{p}_{1,B}, \nu_\varepsilon) \le W_1(\widehat{p}_{1,B}, \nu_B) + W_1(\nu_B, \nu_\varepsilon) \le e^{L_v}\sqrt{\frac{d^\star}{q_{\min}}\mathcal{R}_{\mathrm{FM},B}} + 2\delta_{\mathrm{mix}}. \tag{106}$$

This term propagates through the proof of Theorem 1 without requiring a new geometric argument. Let

$$A_{\mathrm{id}} := \frac{1 + 2C_{\mathrm{id}}}{p_{\min}}, \qquad \mathcal{R}_{\mathrm{FBDM},B} := \mathcal{R}_{\mathrm{FM},B} + \lambda\mathcal{R}_{\mathrm{align}}. \tag{107}$$

Replacing (68) by (106) in (91) yields

$$\Delta_f - 2(h_{\mathrm{aug}} + 2b_{\mathrm{aug}}) \ge g_{\mathrm{aug}} - C_{\mathrm{sep}}\sqrt{\mathcal{R}_{\mathrm{FBDM},B}} - 4e^{-L_v}A_{\mathrm{id}}\delta_{\mathrm{mix}}. \tag{108}$$

Consequently, if

$$C_{\mathrm{sep}}\sqrt{\mathcal{R}_{\mathrm{FBDM},B}} + 4e^{-L_v}A_{\mathrm{id}}\delta_{\mathrm{mix}} \le \frac{g_{\mathrm{aug}}}{2}, \tag{109}$$

the same nearest-centroid argument gives

$$\mathrm{Err}(G_f) \le (1-\sigma) + C_{\mathrm{err}}\sqrt{\mathcal{R}_{\mathrm{FBDM},B}}. \tag{110}$$

Thus balanced assignment recovers the exact-marginal case ($\delta_{\mathrm{mix}} = 0$), while capacitated assignment is controlled by its population mixture-weight mismatch. Capacity constraints alone limit maximum occupancy but do not generally force δ_{mix} to vanish.

D Implementation and Experimental Protocols

D.1 Batch Risk and Training Updates

Fix an epoch e and suppress the capacity schedule's dependence on it. On the fixed dataset $\mathcal{D}_N$, the empirical risk averages over the uniformly sampled minibatch index set $\mathcal{I}$ and the random variables ξ:

$$\widehat{\mathcal{R}}_N(\theta,\psi,\phi) = \mathbb{E}_{\mathcal{I},\xi}\left[\widehat{\mathcal{R}}_B(\theta,\psi,\phi;\mathcal{B}_{\mathcal{I}},\xi)\right], \qquad \mathcal{I}\sim \mathrm{Unif}\left(\binom{[N]}{B}\right). \tag{111}$$

Here, $\binom{[N]}{B} := \{\mathcal{I}\subseteq[N] : |\mathcal{I}| = B\}$ denotes the set of all size-B minibatch index sets, $\mathcal{B}_{\mathcal{I}} = \{\mathbf{x}_i : i\in\mathcal{I}\}$, and $\mathrm{Unif}\left(\binom{[N]}{B}\right)$ is the uniform law over this set. The expectation averages the scalar forward objective. The detached assignment and the scaled backward rule below specify its implemented training updates.

D.2 Two-Timescale Gradient Refinement

Unlike conventional flow matching, whose source law is fixed, FBDM jointly learns the source representations and the velocity field. Encoder updates thus change the states, paths, and target velocities defining the regression problem. For the four reported ResNet-18 benchmarks, we let $\boldsymbol{v}_\phi$ track the current representation geometry before that geometry moves substantially: we use $\rho = 0.75$ to attenuate only the FM gradient entering the encoder and projector, while retaining the full velocity-field and alignment updates. The ImageNet-1K run uses $\rho = 1$ and does not apply this gradient refinement.

For a scale $0 < \rho \le 1$, this is implemented by the forward-identical surrogate

$$\widetilde{\mathbf{z}} = \mathrm{sg}(\mathbf{z}) + \rho(\mathbf{z} - \mathrm{sg}(\mathbf{z})). \tag{112}$$

The surrogate replaces $\mathbf{z}$ in both the FM state and its analytic target; alignment uses the original representation and assignment stays detached. It satisfies $\widetilde{\mathbf{z}} = \mathbf{z}$ in the forward pass and $\partial\widetilde{\mathbf{z}}/\partial\mathbf{z} = \rho\mathbf{I}$. Consequently, for the ResNet-18 objective (18), the training gradients are

$$\nabla^{\mathrm{train}}_{\theta,\psi}\widehat{\mathcal{R}}_B = \rho\,\nabla_{\theta,\psi}\widehat{\mathcal{R}}_{\mathrm{FM},B} + \lambda\,\nabla_{\theta,\psi}\widehat{\mathcal{R}}_{\mathrm{align},B}, \tag{113}$$

$$\nabla^{\mathrm{train}}_{\phi}\widehat{\mathcal{R}}_B = \nabla_{\phi}\widehat{\mathcal{R}}_{\mathrm{FM},B}. \tag{114}$$

Thus ρ changes only the representation-side FM update; it neither changes the forward objective nor weakens velocity regression or view alignment. Algorithm 1 gives the core loop. Main benchmark results use fixed evaluation; online-only diagnostics are explicitly identified in Appendix E.3.

Velocity-energy regularization. The velocity field is auxiliary during pretraining, whereas the encoder is retained downstream. If the field is left unconstrained, it may absorb much of the discrepancy between a representation and its assigned target through a strong transport, reducing the pressure on the encoder to adapt toward the reference geometry. Only for ImageNet-1K, we add velocity-energy regularization to the flow-matching and alignment terms in (16) and (17):

$$\widehat{\mathcal{R}}^{\mathrm{ImageNet}}_B = 1.3\,\widehat{\mathcal{R}}_{\mathrm{FM},B} + 1.3\,\widehat{\mathcal{R}}_{\mathrm{align},B} + 20\,\mathbb{E}_{\mathbf{Z}_T,T}\left[\|\boldsymbol{v}_\phi(\mathbf{Z}_T,T)\|_2^2\right].$$

The energy coefficient is $\beta = 20$ on ImageNet-1K and the energy penalty is not used for the four ResNet-18 benchmarks. This penalty controls the division of optimization burden between representation learning and transport fitting, without removing the transport mechanism. Appendix E.1 (Table 8) compares this velocity-mediated objective with direct regression to the same assigned targets.

D.3 Evaluation and Configuration

Evaluation protocol. The four benchmarks in Table 1 use the frozen 512-dimensional ResNet-18 backbone feature. Their fixed linear protocol trains a classifier for 1000 epochs with probe seeds $\{0, 1, 2\}$; cosine 5-NN is deterministic for a fixed feature bank. The separate ImageNet-1K comparison in Table 2 uses ResNet-50 features after 100 epochs of pretraining with global batch size 512. For FBDM, the linear probe uses all 1,281,167 ImageNet-1K training images and reports top-1 accuracy on all 50,000 validation images (top-5 is also computed). Both splits use `Resize(256)` $\to$ `CenterCrop(224)` $\to$ `ToTensor()`, without input normalization or random augmentation. The ResNet-50 backbone is frozen with batch normalization in evaluation mode; its 2048-dimensional features are taken before the projection head, without using the velocity field or ODE endpoint. A `Linear(2048,1000)` classifier is trained with cross-entropy for 500 epochs using Adam, initial learning rate 0.01 exponentially decayed each epoch to 10^{-6}, weight decay 5×10^{-6}, batch size 1000, and seed 0. Target-center files, source snapshots, and experiment manifests are hash-pinned.

Training hyperparameters. For the dynamic-capacity variant described in Section 3, we write the training-set-specific schedule as $m_{\mathcal{D}_N}(e)$.

Table 5: Dataset-specific assignment capacities used by the reported FBDM runs. Here B is the minibatch size and $m_{\mathcal{D}_N}(e)$ is the number of available slots per center at epoch e.

Dataset	B	K'	$d^\star$	$m_{\mathcal{D}_N}(e)$
CIFAR-10	512	128	64	$5\ (e < 100),\ 7\ (100 \le e < 250),\ 9\ (e \ge 250)$
CIFAR-100	512	160	64	$5\ (e < 125),\ 7\ (125 \le e < 275),\ 9\ (e \ge 275)$
STL-10	512	128	64	$5\ (e < 100),\ 7\ (100 \le e < 250),\ 9\ (e \ge 250)$
Tiny ImageNet	1024	320	96	$5\ (e < 100),\ 7\ (100 \le e < 250),\ 9\ (e \ge 250)$
ImageNet-1K	512	4096	512	4

The two clock families used by the reported FBDM runs are

$$s_{\mathrm{rat},a}(t) = \frac{(1+a)t}{1+at}, \qquad s_{\exp,a}(t) = \frac{1-\exp(-at)}{1-\exp(-a)}. \tag{115}$$

Table 6 records the exact choice for every FBDM dataset entry in Tables 1 and 2. Within each dataset, all reported evaluation entries use the same time protocol.

Table 6: Time reparameterization for the FBDM entries in Tables 1 and 2. Path-uniform sampling draws $S \sim \mathrm{Unif}([0,1])$ and sets $T = s^{-1}(S)$.

Dataset	Reported entries	Clock	Sampling convention
CIFAR-10	Linear, 5-NN	rational, $a = 0.5$	path-uniform
CIFAR-100	Linear, 5-NN	rational, $a = 0.5$	path-uniform
STL-10	Linear, 5-NN	exponential, $a = 1$	path-uniform
Tiny ImageNet	Linear, 5-NN	exponential, $a = 2$	path-uniform
ImageNet-1K	Linear	rational, $a = 0.5$	path-uniform

Table 7: Core configuration for the reported ResNet-18 FBDM results on CIFAR-10, CIFAR-100, STL-10, and Tiny ImageNet.

Component	Setting
Backbone	ResNet-18
Reference geometry	Simplex-spectral
K', $d^\star$, and capacity	See Table 5
Center perturbation	10^{-3}
Velocity time feature	raw t
Flow path	Spherical geodesic
FM time weighting	none ($\gamma = 0$)
Velocity field	4 layers, width 2048, ReLU
Alignment weight	50
FM encoder grad scale	0.75
Optimizer	Adam, lr 3×10^{-5}
Coupled L^2	10^{-4}
Batch size	1024 (Tiny ImageNet); 512 otherwise
Pretraining length	500 epochs
Augmentation	grayscale .1, Gaussian blur .25

The reported ResNet-18 configuration uses RandomResizedCrop with scale $[0.2, 1]$ and aspect ratio $[0.75, 4/3]$, horizontal flip probability .5, ColorJitter $(.4, .4, .4, .1)$ applied with probability .8, grayscale probability .1, and Gaussian blur probability .25. The two views draw independent randomness from the same distribution except in the asymmetric-blur ablation in Table 10.

ImageNet-1K configuration. We use a ResNet-50 backbone with $K' = 4096$ reference components and auxiliary representation dimension $d^\star = 512$. The physical global batch size is $B = 512$. To stabilize capacitated assignment without increasing the physical optimization batch, we aggregate eight consecutive physical batches and assign them jointly over a logical window of

$$W = 8B = 4096$$

images. Capacity is therefore enforced across 4096 images, while parameter updates use physical batches of 512. Each image contributes two augmented views, but their shared consensus is assigned once, so W counts images rather than views. We use per-component capacity $m = 4$, a simplex-spectral reference with perturbation scale $\varepsilon = 10^{-3}$, and a four-layer velocity network of width 2048. The velocity-energy coefficient is $\beta = 20$; the flow-matching and explicit alignment weights are both 1.3, with full FM gradient to the encoder ($\rho = 1$). Training uses LARS with peak learning rate 0.9, a $25{,}020$-step warm-up, and a 100-epoch cosine schedule. We use spherical geodesic paths, path-uniform sampling, and the rational clock with $a = 0.5$.

ImageNet-1K data augmentation. Two views are sampled independently from the same pipeline. Each view uses a 224×224 random resized crop with scale $[0.08, 1.0]$ and aspect ratio $[3/4, 4/3]$, followed by horizontal flipping with probability 0.5. ColorJitter with brightness, contrast, saturation, and hue strengths $(0.8, 0.8, 0.8, 0.2)$ is applied with probability 0.8, followed by random grayscale conversion with probability 0.2. Gaussian blur with a 23×23 kernel and $\sigma \in [0.1, 2.0]$ is applied with probability 0.1. The image is then converted to a tensor without ImageNet mean–standard deviation normalization.

D.4 Matched Training-Cost Measurement

The reported CIFAR-10, CIFAR-100, STL-10, and Tiny ImageNet experiments were each trained on a single NVIDIA V100 GPU. The ImageNet-1K experiments used separate compute setups across runs: one NVIDIA H100, one NVIDIA A100, or four NVIDIA V100 GPUs in parallel. These were not combined in a mixed-GPU run. The DM–FBDM cost comparison below uses a separate matched single-V100 measurement protocol and does not report ImageNet-1K throughput.

The measurements in Table 3 use one Tesla V100-SXM2-16GB with PyTorch 2.2.1 and CUDA 11.8. DM and FBDM both use a ResNet-18 backbone, batch size 512, two views, four data-loading workers, and full precision. Training-loop time is the median after an initial warm-up pass; memory is the

maximum peak allocation and reservation over five measured passes. Evaluation and model-saving I/O are excluded. DM uses five critic updates per encoder update, whereas FBDM uses start-point alignment for each dataset. The comparison therefore measures the native training implementations rather than architecture-independent complexity. Table 3 covers only CIFAR-10, CIFAR-100, and STL-10, the datasets with paired DM and FBDM timing and memory measurements; it does not report a cost comparison for Tiny ImageNet or ImageNet-1K.

E Ablation Studies

E.1 Direct-Target Regression without Flow Matching

To test whether direct matching to the assigned reference targets explains the ImageNet-1K result, we disable the velocity field and compare two regression geometries. Using the normalized representations $\mathbf{z}_j^{(n)}$ from (1) and their shared target $\mathbf{r}^{(n)}$ from (8), define

$$\widehat{\mathcal{R}}_{\mathrm{chord},B} = \frac{1}{2Bd^\star} \sum_{n=1}^{B} \sum_{j=1}^{2} \|\mathbf{z}_j^{(n)} - \mathbf{r}^{(n)}\|_2^2, \tag{116}$$

$$\widehat{\mathcal{R}}_{\mathrm{geo},B} = \frac{1}{2Bd^\star} \sum_{n=1}^{B} \sum_{j=1}^{2} \arccos^2\left((\mathbf{z}_j^{(n)})^\top \mathbf{r}^{(n)}\right). \tag{117}$$

The first uses Euclidean chord distance on the sphere; the second uses the geodesic angle. Both controls retain the explicit two-view alignment term (17) and otherwise use the same assignment, encoder and projector, training procedure, and frozen-backbone evaluation as the reported ImageNet-1K run. Neither trains a velocity field or uses its energy penalty.

Table 8: ImageNet-1K frozen-backbone linear-evaluation top-1 accuracy (%). The direct-regression controls differ only in their spherical distance; both disable flow matching.

Pretraining objective	Top-1
FBDM with velocity energy ($\beta = 20$)	61.41
Direct chord regression, (116)	58.02
Direct geodesic regression, (117)	58.75

Geodesic regression improves on chord regression by 0.73 percentage points, showing that the choice of direct-regression geometry matters. Both direct controls remain below the complete velocity-mediated FBDM result under this evaluation protocol (Table 8).

E.2 Spherical versus Linear Transport Paths

Table 9 compares the spherical-path FBDM results reported in Table 1 with the direct linear alternative to (9), $\boldsymbol{\gamma}^{\mathrm{lin}}_{\mathbf{z},\mathbf{r}}(s) = (1-s)\mathbf{z} + s\mathbf{r}$.

Table 9: Reported FBDM top-1 accuracy (%) with the spherical path used in Table 1 and with direct linear interpolation. Linear denotes frozen-representation linear evaluation; 5-NN uses $k = 5$.

	Spherical path		Linear path	
Dataset	Linear	5-NN	Linear	5-NN
CIFAR-10	92.37	89.68	90.50	88.47
CIFAR-100	66.59	56.74	65.01	55.49
STL-10	89.79	86.23	87.94	83.85
Tiny ImageNet	48.55	33.01	47.47	32.65

E.3 Alignment and Optimization Diagnostics

We test whether shared-target flow matching can replace explicit view alignment. In a matched 1000-epoch diagnostic that retained the complete FM branch but changed only λ from 50 to 0, the final online linear/5-NN accuracies were $80.85\%/75.02\%$ and never exceeded $80.86\%/75.19\%$. Thus the shared endpoint does not substitute for explicit $\mathbf{z}_0$ alignment. Moving explicit alignment to the transported $\mathbf{z}_1$ was also ineffective, peaking online at $78.96\%/72.80\%$. Both diagnostics use online evaluation rather than the fixed protocol in Appendix D.3, so they are excluded from the fixed comparisons below.

Table 10 gives matched fixed-evaluation comparisons on CIFAR-10. Scaling the FM gradient reaching the encoder to 0.75 and removing the long optimizer warmup improve both metrics. Asymmetric two-view blur mainly improves linear accuracy, whereas time weighting with $\gamma = 0.5$ mainly benefits 5-NN.

Table 10: Strict fixed-1000 paired ablations on CIFAR-10. Each row compares a variant with its matched control; Δ is variant minus control.

Factor	Variant / control	Δ Linear	Δ 5-NN
FM encoder gradient	$0.75/1.00$	$+0.860$	$+0.57$
LR warmup steps	$0/500$	$+0.330$	$+0.56$
View blur (p_1, p_2)	$(.20, .30)/(.25, .25)$	$+0.420$	$+0.08$
Time-loss exponent	$0.5/0$	$+0.033$	$+0.89$

Other optimization, augmentation, assignment, flow, sampling, and center variants gave no stable fixed improvement.

F Extended Related Work

Collapse prevention in SSL. Contrastive learning uses negative pairs and uniformity pressure to retain information across the representation space (Chen et al., 2020; Wang & Isola, 2020). Non-contrastive methods rely on predictor asymmetry, target networks, or stop-gradient mechanisms (Grill et al., 2020; Chen & He, 2021). Redundancy-reduction and whitening methods instead constrain second-order statistics (Zbontar et al., 2021; Ermolov et al., 2021; Bardes et al., 2022; Weng et al., 2022). Other approaches use swapped cluster assignments (Caron et al., 2020), self-distillation (Caron et al., 2021), or orthogonality regularization (He et al., 2024). These approaches are strong, but the resulting global representation geometry is an indirect consequence of the anti-collapse mechanism.

Explicit distribution matching. DM replaces defensive regularization with a prescribed reference measure and optimizes a Mallows/Wasserstein discrepancy through a critic (Jiao et al., 2026). FBDM retains the explicit target but replaces the adversarial estimator with a paired transport regression. This distinction is both conceptual and computational: the critic asks whether two distributions can be distinguished, whereas FBDM learns a velocity field for an explicit coupling between them.

Flow matching and geometric transport. Neural ODEs describe continuous transformations through learned vector fields (Chen et al., 2018). Flow matching avoids simulation during training by regressing analytically available conditional velocities (Lipman et al., 2023). Riemannian flow matching extends this construction to manifolds through geodesic conditional paths and gives closed-form targets on simple geometries (Chen & Lipman, 2024). We use its spherical great-circle construction so both interpolants and target velocities respect representation geometry. Unlike generative flow matching, our goal is not sample generation: the learned vector field is a training-time device that shapes the encoder and is discarded after pretraining.

Theory of representation and flow learning. Theoretical accounts of SSL relate representation quality to latent-class transfer, hyperspherical alignment and uniformity, spectral structure, linear transferability, generalization, adversarial robustness, and explicit Wasserstein distribution matching

(Saunshi et al., 2019; Wang & Isola, 2020; HaoChen et al., 2021; 2022; Huang et al., 2023; Duan et al., 2025; Jiao et al., 2026). Separately, flow-matching theory establishes distributional convergence for learned velocity fields, and closely related analyses of probability-flow ODEs characterize convergence in Wasserstein or total-variation distance, numerical discretization error, minimax optimality, and adaptation to intrinsic dimension (Fukumizu et al., 2025; Gao & Zhu, 2025; Huang et al., 2025; Cai & Li, 2025; Li et al., 2024; Tang & Yan, 2025). These results explain representation learning or generative transport in their respective settings, but do not explain why FBDM's auxiliary flow-matching field improves the downstream utility of the encoder. Our analysis targets precisely this missing connection.

Simplex and frame geometry. Simplex equiangular tight frames appear in the neural-collapse literature as maximally separated class means under suitable dimensionality (Papyan et al., 2020). The orthogonal reference construction used in DM requires $K' \leq d^\star$. FBDM instead derives a label-free finite prior from a low-rank approximation of the ideal simplex Gram matrix, extending explicit reference design to the practically useful regime $K' > d^\star$.