DATA SELECTION FOR CONDITIONAL FLOW MATCHING VIA STATE–VELOCITY GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Data selection for Conditional Flow Matching (CFM) asks a basic question: *when can one training sample replace another without changing the state-dependent velocity supervision used to learn the flow?* As Flow Matching scales to larger datasets, full-data training becomes increasingly expensive. However, existing data selection methods typically define sample redundancy through static representations or sample-level scores, so samples that appear redundant may still provide distinct supervision for the learned velocity field. Our key insight is that CFM redundancy should be defined by the *state–velocity geometry* of its supervision, with endpoint distance describing zero-order prescribed supervision and local velocity response capturing first-order sensitivity to state changes. Guided by this view, we introduce **State-Jac**, a lightweight one-shot selector that combines endpoint coverage with proxy-based local-response sketches while leaving final CFM training unchanged. Across balanced, long-tailed, and larger-scale paired benchmarks, State-Jac achieves consistent gains in generation quality and rollout fidelity. On CIFAR-10 and ImageNet-100 at a 20% data budget, State-Jac achieves a best-run FID reduction of 9.2% relative to random selection. On ImageNet-100, its total GPU cost is 24.61% of that of full-data training.

1 INTRODUCTION

Flow Matching (FM) and related interpolant-based formulations learn continuous-time vector fields that transport a source distribution toward data, providing a concise and scalable paradigm for generative modeling (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2025; Ma et al., 2024). Conditional Flow Matching (CFM) provides a tractable training objective by constructing sample-wise state–velocity supervision along conditional paths. As model and dataset scales grow, however, training on the full dataset becomes increasingly expensive. Data selection offers a direct way to reduce this cost by retaining only a representative subset, and substantial redundancy has indeed been observed in generative datasets (Sinha et al., 2020; Briq et al., 2024; Li et al., 2025; Briq et al., 2026b). This raises a basic question for CFM: *when can one training sample replace another without changing the state-dependent velocity supervision used to learn the flow?*

The difficulty is that conventional notions of data redundancy are not tied to the supervision that CFM actually uses. Existing selection methods typically characterize samples through gradient or training statistics, latent representations, scalar difficulty scores, or proxy-model signals (Mirzasoleiman et al., 2020; Paul et al., 2021; Xia et al., 2022; Coleman et al., 2019; Zheng et al., 2025). Such criteria can be highly effective for generic data selection, but they do not directly characterize the state-dependent velocity supervision induced by a CFM training sample. Consequently, two samples that appear redundant in a conventional feature or score space may still provide distinct supervision for the learned vector field. The central question is therefore not merely *which samples are important*, but *what geometry makes two samples interchangeable for CFM?*

We answer this question by deriving a *state–velocity geometry* from CFM supervision and generation dynamics. A training sample determines prescribed states along its conditional path and the velocities supervised at those states. For Rectified Flow, under a shared source coupling, endpoint distance exactly characterizes the joint zero-order discrepancy of prescribed states and target velocities. Yet generation is closed-loop: the learned field is queried on states produced by its own trajectory, so deviations from the prescribed path can change subsequent velocities. This motivates a complementary

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.

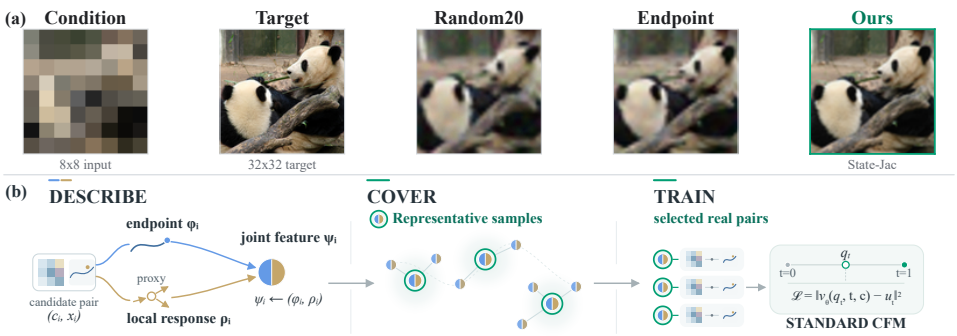


Figure 1: **State-Jac for data selection in Conditional Flow Matching. (a)** An illustrative paired-restoration example at a fixed data budget compares Random20, endpoint-only selection, and State-Jac under the same condition. **(b)** State-Jac describes each sample by its endpoint and local response, covers the resulting joint state–velocity geometry with real representatives, and trains the final CFM on the selected subset using the original objective.

first-order geometry describing how the velocity field responds to changes in state. We show that state-Jacobian discrepancy captures this local sensitivity and can be estimated through shared finite perturbation responses. Together, endpoint distance and local velocity response define two complementary orders of the state–velocity geometry. As summarized in Figure 1, we instantiate this geometry as **State-Jac**, a lightweight one-shot selector that combines endpoint features with proxy-based finite local-response features, covers their joint geometry with real training examples, and trains the final model uniformly on the selected subset with the original CFM objective. State-Jac therefore changes *which samples are used for training*, not *how CFM is trained*.

We evaluate State-Jac in paired instance-conditioned restoration, a controlled setting that allows us to measure generation quality together with condition-matched field and rollout behavior. Across balanced, long-tailed, pixel-space, and latent CFM settings, State-Jac improves generation quality and rollout fidelity, while a controlled cluster-mass intervention shows that better reference-path agreement can coexist with worse own-rollout behavior. On ImageNet-100 at a 20% data budget, State-Jac improves FID from 16.421 ± 0.055 for Random20 to 15.274 ± 0.040 across five final-training seeds. Its total GPU cost is 24.61% of that of full-data training. Code is available in the supplementary material.

Contributions. Our contributions are threefold. **(i) State–velocity geometry.** We derive a CFM-native geometry of sample interchangeability from prescribed supervision and closed-loop dynamics: endpoint distance characterizes zero-order prescribed supervision in Rectified Flow, while local velocity response provides the complementary first-order geometry. **(ii) State-Jac.** We turn this geometry into a lightweight one-shot selector that combines endpoint coverage with proxy-based finite local-response sketches, selects real examples, and leaves final CFM training unchanged. **(iii) Mechanistic and empirical evidence.** Across balanced, long-tailed, and larger-scale paired settings, we show distinct zero- and first-order roles, a controlled counterexample to the sufficiency of reference-path agreement, and consistent gains with substantial end-to-end compute savings.

2 Related Work

Data selection and proxy-guided curation. Data-selection methods estimate sample utility from gradient or curvature information (Mirzasoleiman et al., 2020; Killamsetty et al., 2021; Pooladzandi et al., 2022; Yang et al., 2023), training dynamics or difficulty (Paul et al., 2021; Xia et al., 2022), or coverage in representation spaces. Because signals from the target model can be expensive to obtain, smaller proxy models are widely used to guide data curation (Coleman et al., 2019; Zheng et al., 2025). Recent work further shows that reliable proxy-based curation depends on whether relative comparisons learned from small training runs transfer to the target model scale (Wang et al., 2026a). State-Jac follows this principle at the level of sample geometry: the proxy need only preserve relative relationships between samples, rather than reproduce the final model pointwise.

Data-efficient generative and Flow Matching training. Data-efficient generative training has been studied for GANs and diffusion models through limited-data augmentation, coresets, pruning, clustering, and reweighting (Karras et al., 2020; Zhao et al., 2020; Sinha et al., 2020; Briq et al., 2024; Li et al., 2025). More recently, Briq et al. (2026b) show that Flow Matching can remain surprisingly stable under substantial dataset pruning, providing direct evidence of substantial redundancy in FM training data. Coreset-based work also constructs data-informed source distributions for hierarchical Rectified Flow (Wang et al., 2026b). State-Jac instead leaves the CFM formulation and source distribution unchanged: it selects real examples from a fixed candidate pool and trains the final model uniformly on the selected subset.

Flow dynamics and state–velocity geometry. Recent analyses show that Flow Matching dynamics depend on more than pointwise regression. Wan et al. (2025) connect FM trajectories to underlying data geometry, while Huang et al. (2026) show that small CFM regression error alone need not guarantee accurate probability-path evolution. Recent attribution work further shows that training-sample influence depends on how velocity-field changes propagate along generated trajectories (Briq et al., 2026a). These works analyze flow geometry, path evolution, or training-data attribution; State-Jac instead turns this dynamical structure into a pairwise geometry between candidate samples for pre-training subset selection. Specifically, State-Jac combines zero-order prescribed supervision, characterized by endpoint distance, with first-order local response measured through $D_q v$, where the derivative is taken with respect to the generative state rather than model parameters. Randomized directional probing provides an efficient way to access such implicit Jacobian information without materializing full Jacobians (Hutchinson, 1989; Avron & Toledo, 2011; Grathwohl et al., 2018; Bai et al., 2021; Fournier et al., 2023).

3 State–Velocity Geometry for CFM Sample Interchangeability

What geometry makes two CFM training samples interchangeable? We approach this question through two complementary orders of *state–velocity geometry*. Endpoint distance characterizes zero-order prescribed supervision; closed-loop generation reveals what this geometry leaves unspecified; and local velocity response supplies the complementary first-order geometry (Figure 2). Throughout, x denotes a target, z a source realization, c a condition, and $q \in \mathbb{R}^d$ a generative state; D_q denotes differentiation with respect to q.

3.1 Zero-Order Supervision Geometry

Here, "zero-order" refers to prescribed states and target velocities, before considering the velocity field's response to state perturbations. For straight Rectified-Flow paths (Liu et al., 2022), shared source noise and time make endpoint displacement determine both prescribed-state and target-velocity differences (Figure 2(a)). Averaging squared discrepancies yields the identity below.

Proposition 1 (Zero-order Rectified-Flow equivalence)**.** *Let $x_i, x_j, z \in \mathbb{R}^d$. For $x \in \{x_i, x_j\}$, define the prescribed state and target velocity using the same source realization z and time t:*

$$Q_t(x; z) = (1-t)z + tx, \qquad U(x; z) = x - z, \qquad t \sim \mathrm{Unif}[0, 1]. \tag{1}$$

Their expected joint squared discrepancy is exactly proportional to the squared endpoint distance:

$$\boxed{\mathbb{E}_t\left[\|Q_t(x_i; z) - Q_t(x_j; z)\|_2^2 + \|U(x_i; z) - U(x_j; z)\|_2^2\right] = \frac{4}{3}\|x_i - x_j\|_2^2 .} \tag{2}$$

Indeed, the shared source cancels from both differences, leaving $t(x_i - x_j)$ and $x_i - x_j$; averaging uses $\mathbb{E}_t[t^2] = 1/3$. *Endpoint distance therefore gives an exact zero-order geometry of prescribed state–velocity supervision under this coupling.* Sharing (z, t) isolates sample-induced variation from source and time variation.

The same argument extends to linear paths $Q_t = \alpha(t)z + \beta(t)x$, yielding $C_\beta \|x_i - x_j\|_2^2$ with $C_\beta = \mathbb{E}_t[\beta(t)^2 + \dot{\beta}(t)^2]$; see Appendix A.1. For paired restoration, $c_i = g(x_i)$ and, if g is L_g-Lipschitz, $\|c_i - c_j\|_2 \le L_g \|x_i - x_j\|_2$. Thus endpoint proximity also controls raw-condition proximity in our paired setting, while arbitrary condition-space variation lies outside this identity.

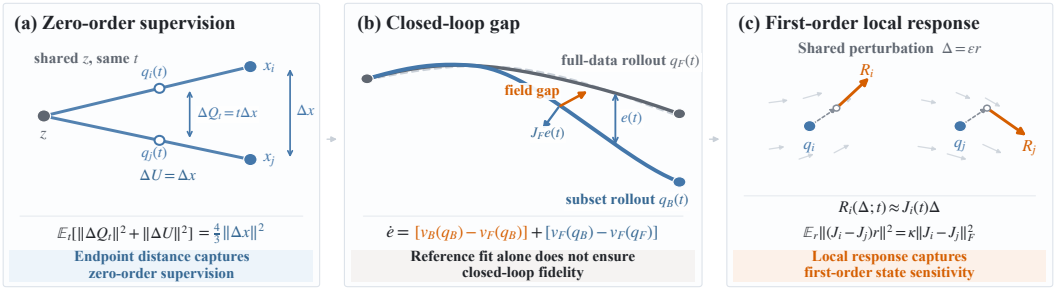


Figure 2: **From prescribed supervision to local-response geometry.** **(a)** Under shared source noise, endpoint distance exactly characterizes zero-order state–velocity supervision in Rectified Flow. **(b)** Generation follows model-dependent rollouts rather than prescribed paths; field mismatch and state-error propagation create a closed-loop gap. **(c)** Shared local perturbations reveal first-order response geometry, where sample similarity depends on how the velocity field reacts to state changes.

3.2 The Closed-Loop Gap

Proposition 1 characterizes similarity on the *prescribed* path. Generation instead evaluates the learned field on states produced by its own rollout (Figure 2(b)). Let v_F and v_B denote fields trained on the full dataset and a selected subset, and $q_F(t), q_B(t)$ their rollout trajectories under condition c:

$$\dot{q}_F = v_F(q_F, t, c), \qquad \dot{q}_B = v_B(q_B, t, c). \tag{3}$$

For trajectory error $e(t) = q_B(t) - q_F(t)$,

$$\dot{e}(t) = \underbrace{v_B(q_B, t, c) - v_F(q_B, t, c)}_{\text{field discrepancy}} + \underbrace{v_F(q_B, t, c) - v_F(q_F, t, c)}_{\text{state-error propagation}}. \tag{4}$$

The first term measures field mismatch at the same visited state; the second captures how accumulated state error changes subsequent velocity. *Thus agreement on prescribed paths alone need not determine closed-loop behavior.* If both trajectories share their initial state and v_F is Lipschitz in state on a region containing both trajectories, with an integrable time-dependent coefficient $L(t)$, Grönwall's inequality gives

$$\boxed{\|e(1)\|_2 \le \int_0^1 \exp\left(\int_t^1 L(s)\,ds\right) \delta_B(t)\,dt,} \tag{5}$$

where $\delta_B(t) = \|v_B(q_B(t), t, c) - v_F(q_B(t), t, c)\|_2$ is the same-state field discrepancy along q_B. The derivation is given in Appendix A.2.

The propagation term exposes a complementary local quantity. Under locally bounded second state derivatives,

$$v_F(q + e, t, c) - v_F(q, t, c) = J_F(q, t, c)e + O(\|e\|_2^2), \qquad J_F := D_q v_F. \tag{6}$$

Thus J_F governs, to first order, how a state deviation changes subsequent velocity. This motivates comparing samples through both prescribed supervision and the field's response to the same local state change.

3.3 First-Order Local-Response Geometry

We use the full-data field $v = v_F$ to define an *ideal* first-order geometry at reference states $q_i(t) = Q_t(x_i; z_i)$. Section 4 will replace this unavailable field with a lightweight proxy. Unlike the zero-order coupling, source realizations z_i need not be shared: here the shared quantity is the state perturbation (Figure 2(c)).

Proposition 2 (First-order local-response equivalence)**.** *Fix a time t, states $q_i(t), q_j(t) \in \mathbb{R}^d$, and conditions c_i, c_j. For each $a \in \{i, j\}$, let $v(\cdot, t, c_a)$ be a C^3 velocity field with bounded second and third state derivatives on a neighborhood containing the closed radius-h ball around $q_a(t)$, for some $h > 0$. Define its local response and state Jacobian by*

$$\mathcal{R}_a(\Delta; t) := v(q_a(t) + \Delta, t, c_a) - v(q_a(t), t, c_a), \qquad J_a(t) := D_q v(q_a(t), t, c_a). \tag{7}$$

Let the shared perturbation $\Delta \in \mathbb{R}^d$ be centrally symmetric about zero, with $\mathbb{E}[\Delta\Delta^\top] = \sigma^2 I$ and $\|\Delta\|_2 \le h$ almost surely. Then its expected squared response discrepancy satisfies

$$\boxed{\mathbb{E}_\Delta \|\mathcal{R}_i(\Delta;t) - \mathcal{R}_j(\Delta;t)\|_2^2 = \sigma^2 \|J_i(t) - J_j(t)\|_F^2 + O\Big(\mathbb{E}_\Delta \|\Delta\|_2^4\Big).} \tag{8}$$

Locally, $\mathcal{R}_i(\Delta;t) = J_i(t)\Delta + O(\|\Delta\|_2^2)$; central symmetry cancels the cubic cross term in the expected squared discrepancy. The proof is given in Appendix A.3. *This complements Proposition 1: endpoint distance compares prescribed supervision, whereas Jacobian discrepancy describes the leading-order difference in local responses.* Averaging over the training-time distribution gives the ideal squared response distance

$$d_{\mathrm{resp}}^2(i,j) := \mathbb{E}_t \|J_i(t) - J_j(t)\|_F^2 . \tag{9}$$

Explicit Jacobians need not be materialized. Standard randomized matrix probing (Avron & Toledo, 2011) expresses their Frobenius discrepancy through shared directional responses.

Proposition 3 (Directional response sketch)**.** *Let $J_i, J_j \in \mathbb{R}^{d\times d}$ be fixed state-Jacobian matrices. If a shared random probe $r \in \mathbb{R}^d$ satisfies $\mathbb{E}[rr^\top] = \kappa I$ for $\kappa > 0$, then*

$$\boxed{\mathbb{E}_r \|(J_i - J_j)r\|_2^2 = \kappa \|J_i - J_j\|_F^2 .} \tag{10}$$

A small number of shared directions provides a finite sketch of this geometry; κ is a global scale absorbed later by calibration. Finite-direction variance and the full derivation are given in Appendix A.4.

Together, endpoint distance and local velocity response form the zero- and first-order components of state–velocity geometry for CFM sample interchangeability. Section 4 makes this geometry practical by replacing the unavailable $J_i r$ with finite local responses from a lightweight proxy; Appendix A.5 analyzes this approximation.

4 State-Jac: Covering State–Velocity Geometry

Section 3 motivates CFM sample interchangeability through zero-order prescribed supervision and first-order local response. **State-Jac** turns this ideal state–velocity geometry into a practical one-shot selector. As summarized in Figure 1(b), given a candidate pool $\mathcal{D} = \{(x_i, c_i)\}_{i=1}^N$ and budget K, State-Jac constructs practical zero- and first-order representations, covers their joint geometry with K real examples, and trains the final CFM uniformly on the selected subset with its original objective unchanged.

4.1 Practical State–Velocity Representation

The zero-order component is directly available from the training samples. Let ϕ_i denote target x_i in the CFM training space. Proposition 1 and its linear-path extension motivate weighting endpoint discrepancy by the path constant C_β, with $C_\beta = 4/3$ for Rectified Flow.

The first-order component is less direct: the ideal response distance in Eq. (9) depends on the full-data field, which is unavailable before subset selection. We therefore train a lightweight proxy v_p once on the candidate training pool using the same CFM objective, and freeze it to estimate relative local-response geometry (Coleman et al., 2019; Zheng et al., 2025; Wang et al., 2026a).

For sample i, fix a source realization z_i, evaluate $q_i(t_k) = Q_{t_k}(x_i; z_i)$ on a shared time grid, and use shared directions $\{r_m\}_{m=1}^M$ to form

$$\widehat{\mathcal{R}}_{p,i}(t_k, r_m) = \frac{v_p(q_i(t_k) + \varepsilon r_m, t_k, c_i) - v_p(q_i(t_k), t_k, c_i)}{\varepsilon}. \tag{11}$$

As $\varepsilon \to 0$, this approaches $J_{p,i}(t_k) r_m$. Thus finite proxy responses provide a computable counterpart of the directional first-order geometry in Proposition 3, without materializing Jacobians.

A fixed shared projection Π compresses these responses, which are concatenated over time–direction pairs and standardized over the candidate pool:

$$\rho_i = \mathrm{Std}_{\mathcal{D}}\Big[\mathrm{concat}_{k,m}\, \Pi\, \widehat{\mathcal{R}}_{p,i}(t_k, r_m)\Big] . \tag{12}$$

The resulting signature ρ_i is a practical surrogate for the first-order local-response geometry. Because endpoint and response features have different numerical scales, we match their average pairwise squared distances over sampled pairs $\mathcal{P}$:

$$\lambda_{\text{match}} = \frac{\mathbb{E}_{(i,j)\sim\mathcal{P}}[C_\beta \, \|\phi_i - \phi_j\|_2^2]}{\mathbb{E}_{(i,j)\sim\mathcal{P}}[\|\rho_i - \rho_j\|_2^2]}. \tag{13}$$

This calibration only aligns the scales of the two theory-derived components; it uses no validation or downstream performance metric. The joint State-Jac representation is

$$\boxed{\psi_i = \begin{bmatrix} \sqrt{C_\beta}\,\phi_i \\ \sqrt{\lambda_{\text{match}}}\,\rho_i \end{bmatrix},} \tag{14}$$

whose two blocks represent the practical zero- and first-order components of state–velocity geometry. Accordingly, squared Euclidean distance in ψ combines endpoint discrepancy with calibrated local-response discrepancy.

4.2 Joint Coverage and Final Training

The resulting geometry measures when one sample can serve as a local substitute for another. To obtain a subset, State-Jac selects K real examples so that every candidate has a nearby representative in this geometry:

$$\boxed{\mathcal{Q}_{\text{SJ}}(\mathcal{B}) = \frac{1}{N}\sum_{i=1}^{N} \min_{b\in\mathcal{B}} \|\psi_i - \psi_b\|_2^2, \qquad |\mathcal{B}| = K.} \tag{15}$$

This gives sample interchangeability a direct operational meaning: each unselected example is represented by a selected example nearby in state–velocity geometry. Restricting representatives to $\mathcal{D}$ preserves valid condition–target pairs. We use scalable approximations to this real-sample quantization objective; details are given in Appendix B.1.

After selection, the proxy, probes, and cached representations are discarded. In the primary State-Jac protocol, the final model is then trained on $\mathcal{B}$ using the original CFM objective with uniform unit weighting. *State-Jac changes which samples are used for training, not how the final CFM is trained.* Selection requires only lightweight proxy training and batched forward evaluations, without materializing full Jacobians or modifying final-model optimization.

5 Experiments

We evaluate the theory-to-method chain through four questions: *(Q1)* does CFM-native state–velocity geometry improve data selection; *(Q2)* do its zero- and first-order components preserve distinct aspects of CFM behavior; *(Q3)* is reference-path agreement sufficient for closed-loop fidelity; and *(Q4)* do the gains persist under imbalance and scale at practical cost?

5.1 Experimental Setup

Why paired restoration? We use conditional super-resolution as a controlled testbed for sample interchangeability. Because each condition $c = g(x)$ is paired with an observed target x, full-data and subset-trained models can be compared under matched conditions and source noise, enabling both generation-quality evaluation and direct field/rollout diagnostics. Paired restoration is a well-established conditional generative setting (Saharia et al., 2022; Albergo et al., 2023). For CIFAR, we use Rectified Flow because its straight path makes the zero-order relation in Proposition 1 exact while retaining the closed-loop dynamics of Section 3.2.

Datasets and metrics. We evaluate CIFAR-10 and CIFAR-100 on $8\times8 \to 32\times32$ super-resolution with a 20% subset budget unless stated otherwise. We additionally study CIFAR-10-LT with imbalance factor 50 at 20% and 40%, and $64\times64 \to 256\times256$ ImageNet-100 restoration with a latent SiT-L/2 CFM model (Ma et al., 2024). We report FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018) for generation quality. We additionally report two rollout metrics: **own**-path error (**Own**),

Table 1: **CIFAR-10 at 20% training budget.** KID, Own, and Drift are multiplied by 10^3 for readability. Lower is better.

Method	Budget	FID ↓	KID ↓	Own ↓	Drift ↓
Full	100%	7.555	4.43	–	–
Random	20%	16.495±.188	12.44±.26	21.343±.177	12.218±.101
DINOv2 K-Medoids	20%	19.287	15.26	22.24	12.85
DINOv2 K-Center	20%	22.694	18.48	23.03	13.43
EL2N-FM High	20%	18.393	**9.53**	26.86	15.16
EL2N-FM Low	20%	34.957	29.39	22.58	13.67
Moderate-DS	20%	16.349±.173	12.19±.23	21.64±.139	12.23±.158
State-Jac	20%	**15.102±.143**	10.88±.20	**20.608±.076**	**11.720±.064**

the mean-squared full/subset velocity discrepancy on the subset model's rollout states; and terminal **drift** (**Drift**), the mean-squared separation between their final states. Lower values indicate closer agreement with the full-data reference. We also report **worst**-class field error (**Worst**), the maximum class-conditional Own. Detailed definitions are given in Appendix B.1.

Baselines and protocol. Baselines span distinct selection principles: Random is information-free; DINOv2 coverage (Oquab et al., 2023) with K-medoids (Jiang & Arias-Castro, 2021) and K-center (Sener & Savarese, 2017) tests representation-space coverage; EL2N-style FM difficulty (Paul et al., 2021) tests scalar difficulty; Moderate-DS (Xia et al., 2022) represents generic data selection; and endpoint-, proxy-residual-, and response-based variants isolate CFM-specific signals, where the residual measures pointwise velocity fitting error on prescribed states. Together, these baselines test generic coverage and difficulty against CFM-specific geometry; State-Jac trains selected examples uniformly with unit loss weights. For CIFAR-10, all primary methods use the same 20% budget, 33,548 training steps, EMA, Euler-100 sampling, and FID-10K. Values with $\pm$ report mean$\pm$std across independent final-training runs; entries without $\pm$ are single runs. State-Jac reuses a fixed selected manifest across runs, whereas Random also resamples the subset, so its variation includes both subset and optimization randomness. Further implementation details are given in Appendix B.1.

5.2 Q1: Does State–Velocity Geometry Improve Data Selection?

Table 1 compares selection principles under the same CIFAR-10 budget. State-Jac achieves the lowest FID, Own, and Drift among the compared subset methods, reducing FID from 16.495 for Random to 15.102 while also improving KID. Relative to Random, DINOv2 coverage worsens both generation quality (FID/KID) and rollout agreement (Own/Drift), whereas EL2N-FM High achieves the best KID but worsens FID, Own, and Drift. *Under the same budget, CFM-native state–velocity geometry therefore provides more consistent joint preservation of generation quality and closed-loop behavior than the generic selection principles tested here.*

5.3 Q2: Do the Zero- and First-Order Components Play Distinct Roles?

Section 3 assigns complementary roles to endpoint and local-response geometry. Table 2 tests this distinction by varying the CFM-specific signals used for selection. On CIFAR-10, Residual achieves the lowest mean FID, while Response-only achieves the lowest mean Own and Drift; State-Jac balances both aspects. Endpoint-only improves over Random, supporting the utility of zero-order coverage. On CIFAR-100, State-Jac has the lowest means across all displayed metrics. Neither component therefore universally dominates; instead, they preserve different aspects of CFM behavior.

Geometry–outcome alignment. We next test this distinction across subset manifests. For subset $\mathcal{B}$, define

$$R_x(\mathcal{B}) = \frac{1}{N}\sum_i \min_{b\in\mathcal{B}} d_x^2(i,b), \qquad R_J(\mathcal{B}) = \frac{1}{N}\sum_i \min_{b\in\mathcal{B}} d_J^2(i,b), \tag{16}$$

where d_x and d_J are endpoint and response distances. We normalize each reported covering distortion by its corresponding Random baseline. Across 30 unique manifests, averaging repeated final-training runs per manifest, endpoint distortion strongly correlates with FID (Spearman $\rho = .89$), whereas response distortion strongly correlates with terminal drift ($\rho = .93$). These are associations rather than deterministic guarantees, but their order-specific alignment matches Section 3.

Table 2: **Zero- and first-order geometry ablations.** End., Res., Resp.: endpoint, proxy residual, and local response. Own, Drift, and Worst are multiplied by 10^3 for readability. Lower is better.

Data	Method	End.	Res.	Resp.	FID ↓	Own ↓	Drift ↓	Worst ↓
CIFAR-10	Endpoint	✓	–	–	15.53	21.20	12.01	–
	Response-only	–	–	✓	15.365±.188	**20.564±.077**	**11.701±.064**	–
	Residual	✓	✓	–	**15.083±.235**	21.232±.111	12.090±.051	–
	State-Jac	✓	–	✓	15.102±.143	20.608±.076	11.720±.064	–
CIFAR-100	Endpoint	✓	–	–	22.730±.314	18.39±.12	10.48±.06	39.43±.16
	Residual	✓	✓	–	21.730±.049	19.12±.26	10.90±.17	36.64±1.31
	State-Jac	✓	–	✓	**21.307±.406**	**18.31±.10**	**10.33±.06**	**35.53±1.22**

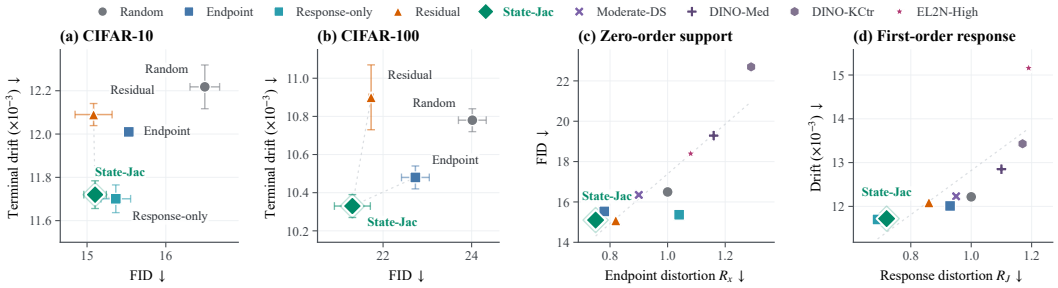


Figure 3: **Complementary roles of zero- and first-order geometry. (a–b)** FID versus terminal drift. **(c)** FID versus endpoint distortion R_x. **(d)** Drift versus response distortion R_J. Lower-left is better.

Theory-to-practice bridge. The theory defines response geometry using the unavailable full-data field, whereas State-Jac estimates it before final training with a lightweight proxy. On CIFAR-10, the tiny proxy preserves substantial full-model response ordering (pairwise Spearman $\rho = .79$, NN-20 overlap .61); increasing proxy capacity raises rank agreement to .87, with only modest further improvement in response distortion. Thus the practical selector need not reproduce the full model pointwise; it only needs to preserve useful relative sample geometry. Appendix B further audits measurement states, response-sketch sensitivity, and selector randomness. *Zero- and first-order geometry capture distinct aspects of CFM behavior: endpoint geometry aligns with generation quality, while local response carries complementary information about closed-loop state sensitivity.*

5.4 Q3: Is Reference-Path Agreement Sufficient for Closed-Loop Fidelity?

We test Section 3.2 with a controlled intervention: holding State-Jac representatives fixed, we replace uniform sampling with cluster-mass sampling, changing only the empirical training measure. Figure 4(a–b) shows a clear reversal: teacher-state and full-path errors decrease by 11.0% and 24.7%, while Own increases by 37.9%, Drift by 51.1%, and FID worsens from 15.102 to 19.426; endpoint representatives show the same pattern. Full results are in Appendix B.5. *This counterexample shows that reference-path agreement is not sufficient for closed-loop fidelity.*

5.5 Q4: Do the Gains Persist under Imbalance and Scale?

Table 3 tests two departures from balanced, small-scale CIFAR training. On CIFAR-10-LT (IF=50), State-Jac improves every reported metric over Random at both 20% and 40% budgets without explicit class balancing; worst-class field error drops from 142.09 to 129.49 and from 55.2 to 46.4, respectively. On ImageNet-100, across five independent final-training seeds, State-Jac improves the mean of all four reported metrics over Random20, including an FID reduction from $16.421 \pm .055$ to $15.274 \pm .040$. The gains therefore extend from balanced pixel-space CIFAR to long-tailed data and a larger latent SiT-L/2 CFM model.

End-to-end cost. Figure 4(c) reports complete ImageNet-100 cost. State-Jac requires 22.889 GPU-hours in total—**24.61% of full-data training compute**—including 20.331 hours of final training, 2.497 hours of proxy training, and 0.061 hours of response extraction; quantization adds 0.308

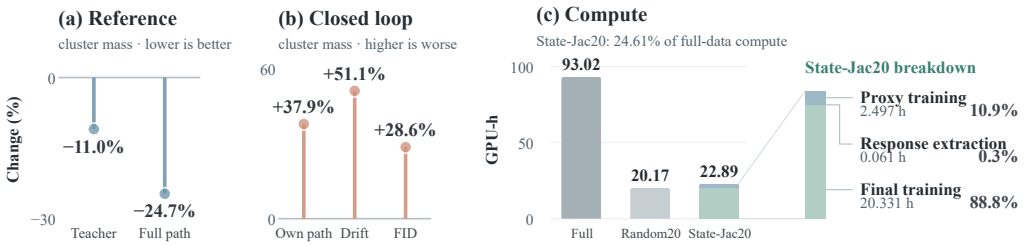


Figure 4: **Closed-loop reversal and efficiency. (a–b)** For fixed State-Jac representatives, cluster-mass sampling improves teacher-state and full-path agreement but worsens own-path error, terminal drift, and FID. **(c)** End-to-end ImageNet-100 GPU cost; the State-Jac bar separates final training, proxy training, and response extraction.

Table 3: **Robustness to imbalance and scale.** CIFAR-10-LT uses imbalance factor 50; ImageNet-100 reports mean±std over five training seeds. KID, Own, Drift, and Worst are multiplied by 10^3 for readability. Lower is better.

Dataset / Budget	Method	FID ↓	KID ↓	Own ↓	Drift ↓	Worst ↓
CIFAR-10-LT (IF=50) / 20%	Random	40.974	32.70	32.52	16.74	142.09
	State-Jac	**38.407**	**29.12**	**31.17**	**16.62**	**129.49**
CIFAR-10-LT (IF=50) / 40%	Random	31.601	24.0	33.7	8.25	55.2
	State-Jac	**30.054**	**22.7**	**33.3**	**8.13**	**46.4**
ImageNet-100 / 100%	Full	12.917	–	–	–	–
ImageNet-100 / 20%	Random20	16.421±.055	4.94±.23	899±39	346±5.6	–
	State-Jac20	**15.274±.040**	**3.07±.18**	**861±31**	**319±7.6**	–

CPU-hours. Thus local-response geometry adds only modest one-time overhead while retaining most of the saving from subset training. Additional analyses are given in Appendix B. *The gains persist under imbalance and larger latent-model scale while the complete State-Jac pipeline remains far below the cost of full-data training.*

6 Conclusion

We studied data selection for Conditional Flow Matching through the question of *sample interchangeability*: when can one training example replace another without changing the supervision relevant to the learned flow? Our analysis answers this through a CFM-native *state–velocity geometry*. For Rectified Flow, endpoint distance exactly characterizes zero-order prescribed supervision, while local velocity response captures the complementary first-order geometry of state sensitivity under closed-loop generation. State-Jac turns these two components into a practical one-shot selector by covering their joint geometry with real training examples, with the final CFM objective unchanged.

Across balanced, long-tailed, and larger-scale paired settings, State-Jac improves generation quality and rollout fidelity with substantial end-to-end compute savings. The experiments further support the distinctions motivated by the theory: zero- and first-order geometry preserve different aspects of CFM behavior, lightweight proxies retain useful relative response geometry, and better reference-path agreement need not imply better closed-loop generation. Taken together, our results suggest that redundancy in CFM is not determined only by where samples lie, but by whether they provide interchangeable supervision for the dynamics of the learned generative field.

Limitations. State-Jac focuses on local first-order response and uses a lightweight proxy to estimate the corresponding sample geometry before final training. The exact endpoint equivalence is established for Rectified Flow and general linear conditional paths, while the closed-loop and local-response analyses rely only on local regularity of the learned field. Empirically, our evaluation focuses on paired instance-conditioned restoration; broader conditioning settings remain an important direction for future work.

Reproducibility Statement

Section 4 specifies the State-Jac selection procedure, while Section 5 reports the principal training and evaluation protocols and end-to-end compute accounting. Appendix A provides complete theoretical derivations, and Appendix B documents proxy approximations, measurement paths, selector and response-sketch stability, controlled interventions, and additional implementation details needed to reproduce the reported analyses.

AI Use Statement

Generative AI tools were used as research-assistance tools for language polishing, code debugging, and figure refinement. All mathematical claims, experimental designs, numerical results, citations, and final manuscript content were reviewed and verified by the authors, who take full responsibility for the submission.

References

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

Michael S Albergo, Mark Goldstein, Nicholas M Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. *arXiv preprint arXiv:2310.03725*, 2023.

Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34, 2011.

Shaojie Bai, Vladlen Koltun, and J Zico Kolter. Stabilizing equilibrium models by jacobian regularization. *arXiv preprint arXiv:2106.14342*, 2021.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Rania Briq, Jiangtao Wang, and Stefan Kesselheim. Data pruning in generative diffusion models. *arXiv preprint arXiv:2411.12523*, 2024.

Rania Briq, Ohad Fried, Michael Kamp, and Stefan Kesselheim. Tracing generated samples to training-data clusters in flow-matching models. *arXiv preprint arXiv:2608.30081*, 2026a.

Rania Briq, Michael Kamp, Ohad Fried, Sarel Cohen, and Stefan Kesselheim. The amazing stability of flow matching. *arXiv preprint arXiv:2604.16079*, 2026b.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019.

Bengt Fornberg. Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of Computation*, 51(184):699–706, 1988.

Louis Fournier, Stéphane Rivaud, Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Can forward gradient match backpropagation? In *International Conference on Machine Learning*, pp. 10249–10264. PMLR, 2023.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Yuhao Huang, Taos Transue, Shih-Hsin Wang, William Feldman, Hong Zhang, and Bao Wang. Improving flow matching by aligning flow divergence. *arXiv preprint arXiv:2602.00869*, 2026.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

He Jiang and Ery Arias-Castro. On the consistency of metric and non-metric k-medoids. In *International Conference on Artificial Intelligence and Statistics*, pp. 2485–2493. PMLR, 2021.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021.

Yize Li, Yihua Zhang, Sijia Liu, and Xue Lin. Pruning then reweighting: Towards data-efficient training of diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607, 2021.

Omead Pooladzandi, David Davini, and Baharan Mirzasoleiman. Adaptive second order coresets for data-efficient machine learning. In *International Conference on Machine Learning*, pp. 17848–17869. PMLR, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Samarth Sinha, Han Zhang, Anirudh Goyal, Yoshua Bengio, Hugo Larochelle, and Augustus Odena. Small-gan: Speeding up gan training using core-sets. In *International Conference on Machine Learning*, pp. 9005–9015. PMLR, 2020.

Zhengchao Wan, Qingsong Wang, Gal Mishne, and Yusu Wang. Elucidating flow matching ode dynamics via data geometry and denoisers. In *Forty-second International Conference on Machine Learning*, 2025.

Jiachen Tianhao Wang, Tong Wu, Kaifeng Lyu, James Y Zou, Dawn Song, Ruoxi Jia, and Prateek Mittal. Can small training runs reliably guide data curation? rethinking proxy-model practice. In *International Conference on Learning Representations*, volume 2026, pp. 122921–122964, 2026a.

Xiao Wang, Zihua She, and Jianxi Su. Coreset-induced conditional velocity flow matching. *arXiv preprint arXiv:2605.12951*, 2026b.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Yu Yang, Hao Kang, and Baharan Mirzasoleiman. Towards sustainable learning: Coresets for data-efficient deep learning. In *International Conference on Machine Learning*, pp. 39314–39330. PMLR, 2023.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in neural information processing systems*, 33:7559–7570, 2020.

Haizhong Zheng, Elisa Tsai, Yifu Lu, Jiachen Sun, Brian Bartoldson, Bhavya Kailkhura, and Atul Prakash. Elfs: Label-free coreset selection with proxy training dynamics. In *International Conference on Learning Representations*, volume 2025, pp. 63391–63410, 2025.

A Additional Theoretical Analysis

This appendix provides the derivations underlying Section 3. We first extend the endpoint identity beyond straight Rectified Flow, then derive the closed-loop trajectory bound and local-response equivalence. We next analyze randomized directional sketches and the finite local-response estimator used by State-Jac, and finally connect the ideal full-model response geometry to its practical proxy-based approximation.

Throughout, vector norms are Euclidean unless otherwise specified, $\|\cdot\|_F$ denotes the Frobenius norm, and expectations are taken over the explicitly indicated random variables.

A.1 Extension to General Linear Probability Paths

Proposition 1 was stated for the straight path of Rectified Flow. The same argument applies to a broader family of linear conditional probability paths commonly considered within Flow Matching (Lipman et al., 2022).

Consider

$$Q_t(x; z) = \alpha(t)z + \beta(t)x, \tag{17}$$

where $\alpha, \beta : [0, 1] \to \mathbb{R}$ are differentiable. The corresponding conditional target velocity is

$$U_t(x; z) = \dot{\alpha}(t)z + \dot{\beta}(t)x. \tag{18}$$

For two samples x_i, x_j evaluated with the same source z and time t, the source-dependent terms cancel:

$$Q_t(x_i; z) - Q_t(x_j; z) = \beta(t)(x_i - x_j), \tag{19}$$

$$U_t(x_i; z) - U_t(x_j; z) = \dot{\beta}(t)(x_i - x_j). \tag{20}$$

Proposition 4 (Endpoint geometry for linear paths)**.** *For the path in Eq.* (17)*,*

$$\boxed{\mathbb{E}_t\left[\|Q_t(x_i; z) - Q_t(x_j; z)\|_2^2 + \|U_t(x_i; z) - U_t(x_j; z)\|_2^2\right] = C_\beta \|x_i - x_j\|_2^2,} \tag{21}$$

where

$$C_\beta := \mathbb{E}_t\left[\beta(t)^2 + \dot{\beta}(t)^2\right]. \tag{22}$$

Proof. Using Eqs. (19) and (20),

$$\begin{aligned}&\|Q_t(x_i; z) - Q_t(x_j; z)\|_2^2 + \|U_t(x_i; z) - U_t(x_j; z)\|_2^2 \\ &\quad = \left(\beta(t)^2 + \dot{\beta}(t)^2\right) \|x_i - x_j\|_2^2 .\end{aligned} \tag{23}$$

Taking expectation over t proves the claim. □

For Rectified Flow, $\alpha(t) = 1 - t$ and $\beta(t) = t$, giving

$$C_\beta = \mathbb{E}_t[t^2 + 1] = \frac{4}{3}, \tag{24}$$

which recovers Proposition 1.

Thus the zero-order interpretation is not specific to the constant $4/3$: for any linear path, endpoint distance remains proportional to the joint discrepancy between prescribed states and target velocities.

Why shared randomness matters. If different sources z_i and z_j were used in the zero-order comparison, Eq. (19) would additionally contain $\alpha(t)(z_i - z_j)$, mixing sample-induced differences with source variation. Shared z and t therefore provide the controlled coupling required for the endpoint identity. This requirement is specific to that comparison: the first-order response equivalence is pointwise in the chosen reference states and instead requires the perturbation to be shared. Accordingly, the practical extractor may use fixed per-sample source realizations while sharing response probes.

A.2 Closed-Loop Trajectory Error Bound

We now derive Eq. (5). Let

$$\dot{q}_F(t) = v_F(q_F(t), t, c), \qquad \dot{q}_B(t) = v_B(q_B(t), t, c), \tag{25}$$

with $q_F(0) = q_B(0)$, and define

$$e(t) := q_B(t) - q_F(t). \tag{26}$$

Adding and subtracting $v_F(q_B(t), t, c)$ gives

$$\begin{aligned}\dot{e}(t) = &\left[v_B(q_B(t), t, c) - v_F(q_B(t), t, c)\right] \\ &+ \left[v_F(q_B(t), t, c) - v_F(q_F(t), t, c)\right].\end{aligned} \tag{27}$$

Assume that along the relevant trajectories, v_F is locally Lipschitz in state:

$$\left\|v_F(q_1, t, c) - v_F(q_2, t, c)\right\|_2 \le L(t) \left\|q_1 - q_2\right\|_2, \tag{28}$$

where $L : [0, 1] \to \mathbb{R}_{\ge 0}$ is integrable. Define

$$\delta_B(t) := \left\|v_B(q_B(t), t, c) - v_F(q_B(t), t, c)\right\|_2. \tag{29}$$

Integrating Eq. (27) and applying the triangle inequality gives

$$\left\|e(t)\right\|_2 \le \int_0^t \delta_B(s)\, ds + \int_0^t L(s) \left\|e(s)\right\|_2 ds. \tag{30}$$

The integral form of Grönwall's inequality therefore yields

$$\left\|e(t)\right\|_2 \le \int_0^t \exp\left(\int_s^t L(u)\, du\right) \delta_B(s)\, ds. \tag{31}$$

Setting $t = 1$ recovers Eq. (5).

Interpretation. The bound separates the field discrepancy created by subset training from the sensitivity of subsequent dynamics to an accumulated state deviation. If v_F is differentiable in q,

$$v_F(q + e, t, c) - v_F(q, t, c) = J_F(q, t, c)e + r(q, e, t, c), \tag{32}$$

where

$$\frac{\left\|r(q, e, t, c)\right\|_2}{\left\|e\right\|_2} \to 0 \qquad \text{as } \left\|e\right\|_2 \to 0. \tag{33}$$

Under twice differentiability of v_F, the remainder is $O(\|e\|_2^2)$, giving Eq. (6).

A.3 Proof of First-Order Local-Response Equivalence

We next prove Proposition 2. Fix a time t and write $q_i := q_i(t)$ and $J_i := J_i(t)$. Recall

$$\mathcal{R}_i(\Delta) = v(q_i + \Delta, t, c_i) - v(q_i, t, c_i). \tag{34}$$

Let H_i denote the second state derivative of the vector-valued field. Under bounded third state derivatives, Taylor's theorem gives

$$\mathcal{R}_i(\Delta) = J_i\Delta + \frac{1}{2}H_i[\Delta, \Delta] + R_i^{(3)}(\Delta), \tag{35}$$

with $\left\|R_i^{(3)}(\Delta)\right\|_2 \le C_i \left\|\Delta\right\|_2^3$ within a sufficiently small neighborhood.

Define

$$A := J_i - J_j, \qquad B := H_i - H_j, \tag{36}$$

and $R^{(3)} := R_i^{(3)} - R_j^{(3)}$. Then

$$\mathcal{R}_i(\Delta) - \mathcal{R}_j(\Delta) = A\Delta + \frac{1}{2}B[\Delta, \Delta] + R^{(3)}(\Delta). \tag{37}$$

The leading squared term satisfies

$$\begin{aligned}\mathbb{E}_\Delta \|A\Delta\|_2^2 &= \operatorname{tr}\left(A^\top A \mathbb{E}[\Delta\Delta^\top]\right) \\ &= \sigma^2 \|A\|_F^2 .\end{aligned} \tag{38}$$

The potentially third-order cross term

$$\langle A\Delta, B[\Delta,\Delta]\rangle \tag{39}$$

is odd under $\Delta \mapsto -\Delta$. Central symmetry therefore gives

$$\mathbb{E}_\Delta\left[\langle A\Delta, B[\Delta,\Delta]\rangle\right] = 0. \tag{40}$$

All remaining terms are fourth order or higher under the stated derivative bounds. Consequently,

$$\boxed{\mathbb{E}_\Delta \|\mathcal{R}_i(\Delta;t) - \mathcal{R}_j(\Delta;t)\|_2^2 = \sigma^2 \|J_i(t) - J_j(t)\|_F^2 + O\Big(\mathbb{E}\,\|\Delta\|_2^4\Big).} \tag{41}$$

This proves Proposition 2.

Why central symmetry matters. Without central symmetry, the third-order cross term need not vanish and the leading correction can be $O(\mathbb{E}\,\|\Delta\|_2^3)$. Symmetric probes eliminate this odd-order contribution. Importantly, this is a property of the perturbation distribution, not of a particular finite-difference stencil; the shared Rademacher probes used by State-Jac satisfy this symmetry.

A.4 Directional Sketch and Finite-Sample Approximation

Let

$$A := J_i - J_j. \tag{42}$$

For a random direction r satisfying $\mathbb{E}[rr^\top] = \kappa I$,

$$\begin{aligned}\mathbb{E}_r \|Ar\|_2^2 &= \operatorname{tr}\left(A^\top A \mathbb{E}[rr^\top]\right) \\ &= \kappa \|A\|_F^2 ,\end{aligned} \tag{43}$$

which proves Proposition 3.

This covers Gaussian and Rademacher probes; for unnormalized standard Gaussian or Rademacher directions, $\kappa = 1$, while normalizing a Rademacher vector by $1/\sqrt{d}$ gives $\kappa = 1/d$. Since State-Jac subsequently calibrates the response scale relative to the endpoint component, this global factor does not affect relative geometry.

For M independent directions, define

$$\widehat{D}_M(A) := \frac{1}{M}\sum_{m=1}^{M} \|Ar_m\|_2^2 . \tag{44}$$

Then

$$\mathbb{E}[\widehat{D}_M(A)] = \kappa \|A\|_F^2 , \tag{45}$$

and

$$\operatorname{Var}[\widehat{D}_M(A)] = \frac{1}{M}\operatorname{Var}[\|Ar\|_2^2]. \tag{46}$$

Thus Monte Carlo variance decreases at the standard $1/M$ rate.

If $\|r\|_2 \le R$ almost surely and $\|A\|_2 \le S$, then $0 \le \|Ar\|_2^2 \le S^2R^2$, and Hoeffding's inequality gives

$$\Pr\Big(\Big|\widehat{D}_M(A) - \kappa\|A\|_F^2\Big| \ge \tau\Big) \le 2\exp\Big(-\frac{2M\tau^2}{S^4R^4}\Big). \tag{47}$$

Time averaging. If t_k are sampled from the CFM time distribution, then

$$\frac{1}{TM}\sum_{k=1}^{T}\sum_{m=1}^{M}\|(J_i(t_k)-J_j(t_k))r_m\|_2^2 \tag{48}$$

is unbiased, up to κ, for $\mathbb{E}_t\|J_i(t)-J_j(t)\|_F^2$. With a deterministic time grid, it instead provides the corresponding quadrature approximation.

A.5 Finite Local Responses and First-Order Geometry

For the proxy field, define

$$J_{p,i}(t) := D_q v_p(q_i(t),t,c_i), \qquad \mathcal{R}_i^p(\Delta;t) := v_p(q_i(t)+\Delta,t,c_i)-v_p(q_i(t),t,c_i). \tag{49}$$

The practical CIFAR extractor does not explicitly form $J_{p,i}(t)r$. Instead it evaluates the normalized finite local response

$$\widehat{\mathcal{R}}_{i,\varepsilon}(r;t) := \frac{v_p(q_i(t)+\varepsilon r,t,c_i)-v_p(q_i(t),t,c_i)}{\varepsilon} = \frac{\mathcal{R}_i^p(\varepsilon r;t)}{\varepsilon}. \tag{50}$$

This is the proxy analogue of the local response in Eq. (7). Viewed pointwise as a one-sided directional-derivative approximation,

$$\widehat{\mathcal{R}}_{i,\varepsilon}(r;t) = J_{p,i}(t)r + O(\varepsilon\|r\|_2^2), \tag{51}$$

under bounded second state derivatives (Fornberg, 1988).

For State-Jac, however, the quantity of interest is the *pairwise response geometry*, for which the local-response equivalence gives a stronger statement.

Proposition 5 (Finite-response geometry)**.** *Assume bounded second and third state derivatives of v_p in sufficiently small neighborhoods of $q_i(t)$ and $q_j(t)$. Let r be centrally symmetric with $\mathbb{E}[rr^\top]=\kappa I$ and finite fourth moment, and assume $q_i(t)+\varepsilon r$ and $q_j(t)+\varepsilon r$ remain in these neighborhoods almost surely. Then*

$$\boxed{\begin{aligned}&\mathbb{E}_r\left\|\widehat{\mathcal{R}}_{i,\varepsilon}(r;t)-\widehat{\mathcal{R}}_{j,\varepsilon}(r;t)\right\|_2^2\\ &\qquad=\kappa\|J_{p,i}(t)-J_{p,j}(t)\|_F^2+O\left(\varepsilon^2\mathbb{E}\|r\|_2^4\right).\end{aligned}} \tag{52}$$

Proof. Apply the argument of Proposition 2 to the proxy field with $\Delta=\varepsilon r$. Since $\mathbb{E}[\Delta\Delta^\top]=\varepsilon^2\kappa I$,

$$\begin{aligned}&\mathbb{E}_r\left\|\mathcal{R}_i^p(\varepsilon r;t)-\mathcal{R}_j^p(\varepsilon r;t)\right\|_2^2\\ &\qquad=\varepsilon^2\kappa\|J_{p,i}(t)-J_{p,j}(t)\|_F^2+O\left(\varepsilon^4\mathbb{E}\|r\|_2^4\right).\end{aligned} \tag{53}$$

Dividing by ε^2 proves the result. □

Thus the one-sided estimator has $O(\varepsilon)$ pointwise derivative error, yet its expected *squared pairwise response geometry* has no first-order correction under symmetric probes. State-Jac therefore requires stable relative geometry between samples rather than high-accuracy recovery of each individual Jacobian.

For the reported CIFAR experiments we use $\varepsilon=0.02$. Appendix B.4 verifies that halving or doubling this scale leaves the induced response geometry largely unchanged.

A.6 From Full-Model Own-Rollout Response to Proxy Reference Response

The theoretical motivation in Section 3.2 concerns states reached during rollout, whereas practical State-Jac uses a lightweight proxy evaluated on prescribed reference states. We now make this approximation explicit.

For sample i, define

$$J_i^{F,\mathrm{own}}(t) := D_q v_F(q_i^F(t), t, c_i), \tag{54}$$

$$J_i^{F,\mathrm{ref}}(t) := D_q v_F(q_i^{\mathrm{ref}}(t), t, c_i), \tag{55}$$

$$J_i^{p,\mathrm{ref}}(t) := D_q v_p(q_i^{\mathrm{ref}}(t), t, c_i), \tag{56}$$

where $q_i^F(t)$ is the full-model rollout state and $q_i^{\mathrm{ref}}(t)$ the prescribed interpolation state.

Proxy approximation. Assume

$$\left\| J_i^{p,\mathrm{ref}}(t) - J_i^{F,\mathrm{ref}}(t) \right\|_F \le \eta_p(t) \tag{57}$$

on the reference states of interest.

Path approximation. Assume the full-model Jacobian is locally Lipschitz in state:

$$\|D_q v_F(q_1, t, c) - D_q v_F(q_2, t, c)\|_F \le H(t) \|q_1 - q_2\|_2 . \tag{58}$$

Define

$$\tau_i(t) := \left\| q_i^F(t) - q_i^{\mathrm{ref}}(t) \right\|_2 . \tag{59}$$

Then

$$\left\| J_i^{F,\mathrm{own}}(t) - J_i^{F,\mathrm{ref}}(t) \right\|_F \le H(t)\tau_i(t). \tag{60}$$

For pair (i, j), let

$$A_{\mathrm{own}}(t) := J_i^{F,\mathrm{own}}(t) - J_j^{F,\mathrm{own}}(t), \tag{61}$$

$$A_{p,\mathrm{ref}}(t) := J_i^{p,\mathrm{ref}}(t) - J_j^{p,\mathrm{ref}}(t). \tag{62}$$

Proposition 6 (Proxy/reference approximation)**.** *Under Eqs.* (57) *and* (58)*,*

$$\boxed{\|A_{\mathrm{own}}(t) - A_{p,\mathrm{ref}}(t)\|_F \le 2\eta_p(t) + H(t)\left(\tau_i(t) + \tau_j(t)\right).} \tag{63}$$

Consequently,

$$\begin{aligned} &\left| \|A_{\mathrm{own}}(t)\|_F - \|A_{p,\mathrm{ref}}(t)\|_F \right| \\ &\quad \le 2\eta_p(t) + H(t)\left(\tau_i(t) + \tau_j(t)\right). \end{aligned} \tag{64}$$

Proof. Add and subtract the corresponding full-model reference Jacobians, then apply the triangle inequality together with Eqs. (57)–(60). The second statement follows from the reverse triangle inequality. □

For the time-averaged RMS metrics

$$d_{\mathrm{own}}(i, j) := \left(\mathbb{E}_t \|A_{\mathrm{own}}(t)\|_F^2 \right)^{1/2}, \tag{65}$$

$$d_{p,\mathrm{ref}}(i, j) := \left(\mathbb{E}_t \|A_{p,\mathrm{ref}}(t)\|_F^2 \right)^{1/2}, \tag{66}$$

Minkowski's inequality gives

$$\boxed{\begin{aligned} &|d_{\mathrm{own}}(i, j) - d_{p,\mathrm{ref}}(i, j)| \\ &\quad \le \left(\mathbb{E}_t \left[2\eta_p(t) + H(t)(\tau_i(t) + \tau_j(t)) \right]^2 \right)^{1/2}. \end{aligned}} \tag{67}$$

Thus the gap between ideal own-rollout geometry and proxy-reference geometry is controlled by two distinct quantities: proxy error in local Jacobian geometry and the reference-to-rollout state gap. These are tested directly by the proxy and measurement-path diagnostics in Appendix B.

A.7 Combining Proxy, Finite-Response, and Sampling Effects

The preceding results expose the successive approximations made by the practical selector. Define the proxy-reference squared response geometry

$$G_{p,\mathrm{ref}}(i,j) := \mathbb{E}_t \left\| A_{p,\mathrm{ref}}(t) \right\|_F^2 , \tag{68}$$

and its population finite-response counterpart

$$G_{p,\varepsilon}(i,j) := \frac{1}{\kappa} \mathbb{E}_{t,r} \left\| \widehat{\mathcal{R}}_{i,\varepsilon}(r;t) - \widehat{\mathcal{R}}_{j,\varepsilon}(r;t) \right\|_2^2 . \tag{69}$$

Under uniform derivative and moment bounds, Proposition 5 implies

$$G_{p,\varepsilon}(i,j) = G_{p,\mathrm{ref}}(i,j) + O(\varepsilon^2). \tag{70}$$

The remaining approximation replaces the expectation over time and direction by a finite time grid and finitely many shared probes. For random directions, Eq. (46) gives the usual $1/M$ variance decay; a deterministic time grid introduces the corresponding quadrature approximation.

The practical chain can therefore be summarized as

$$\boxed{\begin{aligned} &\text{full own-rollout response} \\ &\quad \longrightarrow \text{full reference response} \\ &\quad \longrightarrow \text{proxy reference response} \\ &\quad \longrightarrow \text{finite local response} \\ &\quad \longrightarrow \text{finite response sketch} \\ &\quad \longrightarrow \text{compact response signature.} \end{aligned}} \tag{71}$$

The first two gaps are controlled by Proposition 6, the finite-response gap by Eq. (70), and directional sampling by Eq. (46). The implementation further applies a fixed shared linear compression and coordinate-wise standardization to obtain the compact response signature. These operations are used to preserve a practical relative geometry, not to reconstruct the full Jacobian.

Accordingly, State-Jac does not require every approximation above to vanish individually; it requires the resulting *relative sample geometry* to remain informative for selection. Appendix B tests precisely this requirement.

A.8 Uniform Representatives and Cluster-Mass Approximation

State-Jac selects actual training examples but trains the final model uniformly over the selected subset. This differs from weighted coreset constructions that explicitly preserve the mass represented by each selected point (Mirzasoleiman et al., 2020). The following elementary result separates geometric quantization error from cluster-mass mismatch.

Let $a_i := \psi_i$ denote the joint State-Jac representation and $b(i) \in \mathcal{B}$ the nearest selected representative to sample i. For each selected representative b_k, define

$$C_k := \{ i : b(i) = b_k \}, \qquad p_k := \frac{|C_k|}{N}. \tag{72}$$

Consider any scalar functional f satisfying

$$|f(a) - f(a')| \le L_f \left\| a - a' \right\|_2 , \qquad |f(a)| \le M_f. \tag{73}$$

Lemma 1 (Uniform-representative approximation)**.** *The discrepancy between the full empirical average and the uniformly weighted representative average satisfies*

$$\boxed{\begin{aligned} &\left| \frac{1}{N} \sum_{i=1}^{N} f(a_i) - \frac{1}{K} \sum_{k=1}^{K} f(a_{b_k}) \right| \\ &\le L_f \frac{1}{N} \sum_{i=1}^{N} \left\| a_i - a_{b(i)} \right\|_2 + M_f \sum_{k=1}^{K} \left| p_k - \frac{1}{K} \right| . \end{aligned}} \tag{74}$$

Proof. Add and subtract the cluster-mass-weighted representative average:

$$\left|\frac{1}{N}\sum_i f(a_i) - \frac{1}{K}\sum_k f(a_{b_k})\right| \\ \leq \left|\frac{1}{N}\sum_i f(a_i) - \sum_k p_k f(a_{b_k})\right| + \left|\sum_k \left(p_k - \frac{1}{K}\right) f(a_{b_k})\right|. \tag{75}$$

The first term is at most

$$\frac{L_f}{N}\sum_i \left\|a_i - a_{b(i)}\right\|_2, \tag{76}$$

and the second is at most

$$M_f \sum_k \left|p_k - \frac{1}{K}\right|. \tag{77}$$

Combining them proves the claim. □

Since

$$\mathcal{Q}_{\mathrm{SJ}}(\mathcal{B}) = \frac{1}{N}\sum_i \left\|a_i - a_{b(i)}\right\|_2^2, \tag{78}$$

Cauchy–Schwarz gives

$$\frac{1}{N}\sum_i \left\|a_i - a_{b(i)}\right\|_2 \leq \sqrt{\mathcal{Q}_{\mathrm{SJ}}(\mathcal{B})}. \tag{79}$$

Hence

$$\boxed{\begin{aligned}&\left|\frac{1}{N}\sum_i f(a_i) - \frac{1}{K}\sum_k f(a_{b_k})\right| \\ &\leq L_f\sqrt{\mathcal{Q}_{\mathrm{SJ}}(\mathcal{B})} + M_f \sum_k \left|p_k - \frac{1}{K}\right|.\end{aligned}} \tag{80}$$

Equation (80) separates *geometric representation error*, captured by State-Jac quantization, from *occupancy imbalance* among selected regions. This result does not prescribe reweighting; it clarifies why geometric coverage and the empirical training measure are distinct design choices.

A.9 Summary of the Theoretical Chain

The theoretical development follows the picture in Figure 2:

$$\boxed{\begin{aligned}\text{endpoint geometry} &\longrightarrow \text{zero-order prescribed supervision} \\ &\xrightarrow{\text{closed-loop gap}} \text{local state sensitivity} \\ &\longrightarrow \text{first-order local response} \\ &\longrightarrow \text{directional / finite-response sketch} \\ &\longrightarrow \text{practical State-Jac geometry.}\end{aligned}} \tag{81}$$

Proposition 1 establishes the zero-order identity; Eq. (5) exposes the dependence on model-visited states; Proposition 2 identifies Jacobian discrepancy as the first-order geometry of local velocity responses; Proposition 3 makes this geometry tractable with shared probes; and Proposition 5 connects the actual finite local-response estimator to the same pairwise geometry. The experiments then test whether these geometric components preserve the corresponding aspects of generation and rollout behavior.

Table 4: **Primary training and evaluation settings.** The proxy is used only for selection and discarded before final training.

Benchmark	Condition → target	Budget	Final model	Main protocol
CIFAR-10	$8^2 \to 32^2$	20%	RF DDPM++	33,548 steps, EMA, Euler-100
CIFAR-100	$8^2 \to 32^2$	20%	RF DDPM++	same pixel-space protocol
CIFAR-10-LT	$8^2 \to 32^2$	20%, 40%	RF DDPM++	IF $= 50$
ImageNet-100	$64^2 \to 256^2$	20%	SiT-L/2	latent CFM, 5 train seeds

B Practical Validation and Additional Experiments

Section 3 defines an ideal first-order response geometry through the state Jacobian of the full-data vector field, whereas State-Jac must construct its selection geometry *before* final training. The practical method therefore replaces this ideal object with a lightweight proxy, prescribed reference states, finitely many shared probes, finite local responses, and a compact response signature. This appendix documents the practical selection protocol and tests these approximations directly.

The practical premise does not require pointwise reproduction of the full model. Instead, we test whether the *relative geometry between samples* is preserved well enough to identify redundant regions. We therefore emphasize rank agreement, neighborhood preservation, covering distortion, selector stability, controlled interventions, and downstream alignment.

B.1 Implementation and Evaluation Protocol

Benchmarks and final training. Table 4 summarizes the primary settings. CIFAR-10 and CIFAR-100 use paired $8\times8 \to 32\times32$ conditional super-resolution with a 20% subset budget. CIFAR-10-LT uses the same pixel-space setting with imbalance factor 50 and 20% or 40% budgets. ImageNet-100 uses 64×64 conditions and 256×256 targets in latent space with a SiT-L/2 final model. Within each benchmark, Full, Random, and curated-subset runs share the same final-training pipeline; selection changes the training manifest rather than the CFM objective.

For CIFAR, the selection proxy uses the same DDPM++ model family with base width 64 and is trained on the full candidate training pool for a fixed budget of 8,000 optimization steps. We use Adam with batch size 128 and learning rate 5×10^{-5}, with linear warmup over the first 5,000 steps. Response features are extracted using the EMA weights at step 8,000, without validation-based early stopping or checkpoint selection. Generation quality is evaluated with FID-10K using Euler-100 sampling. The same final-training conventions are used for the related pixel-space benchmarks unless stated otherwise.

Response extraction, compression, and source coupling. The shared-source construction in Proposition 1 is a controlled coupling used to derive the zero-order identity; it is not required by the first-order response comparison. The practical extractor instead assigns each example a fixed deterministic source realization z_i,

$$q_i(t_k) = Q_{t_k}(x_i; z_i), \tag{82}$$

while sharing the time grid and response directions across samples.

For the reported CIFAR experiments,

$$t_k \in \{0.2, 0.4, 0.6, 0.8\}, \qquad M = 2, \qquad \varepsilon = 0.02, \tag{83}$$

with two shared Rademacher state directions. We evaluate the one-sided normalized local response

$$\widehat{\mathcal{R}}_{p,i}(t_k, r_m) = \frac{v_p(q_i(t_k) + \varepsilon r_m, t_k, c_i) - v_p(q_i(t_k), t_k, c_i)}{\varepsilon}, \tag{84}$$

corresponding to Eq. (11).

Each 3072-dimensional CIFAR velocity response is mapped to 64 dimensions using a fixed Rademacher projection shared across the candidate pool. The 4×2 time–direction blocks are

Algorithm 1 STATE-JAC One-Shot Data Selection

Require: Candidate pool $\mathcal{D} = \{(x_i, c_i)\}_{i=1}^{N}$; budget K; times $\{t_k\}_{k=1}^{T}$; shared directions $\{r_m\}_{m=1}^{M}$; response scale ε
Ensure: Selected subset $\mathcal{B}$
Train lightweight CFM proxy v_p on $\mathcal{D}$
Fix one deterministic source realization z_i per sample
for $i = 1, \ldots, N$ **do**
    Set endpoint feature ϕ_i to the training-space target
    for $k = 1, \ldots, T$ and $m = 1, \ldots, M$ **do**
        $q_i(t_k) \leftarrow Q_{t_k}(x_i; z_i)$
        Compute $\widehat{\mathcal{R}}_{p,i}(t_k, r_m) \leftarrow [v_p(q_i(t_k) + \varepsilon r_m, t_k, c_i) - v_p(q_i(t_k), t_k, c_i)]/\varepsilon$
        Project response block with the shared map Π
    end for
    Concatenate projected blocks to obtain raw signature $\tilde{\rho}_i$
end for
Standardize coordinates of $\{\tilde{\rho}_i\}_{i=1}^{N}$ over $\mathcal{D}$ to obtain $\{\rho_i\}_{i=1}^{N}$
Estimate λ_{match} from training-pair geometry
Form $\psi_i = [\sqrt{C_\beta}\,\phi_i; \sqrt{\lambda_{\text{match}}}\,\rho_i]$
$\mathcal{B} \leftarrow$ REALSAMPLEQUANTIZE$(\{\psi_i\}_{i=1}^{N}, K)$
return $\mathcal{B}$

concatenated into a 512-dimensional signature and standardized coordinate-wise over the training pool. Thus the practical representation is designed to preserve relative response geometry rather than reconstruct the full Jacobian.

The scale calibration in Eq. (13) is estimated from 200,000 uniformly sampled training pairs and does not use validation FID or another downstream metric. Endpoint and response features are extracted once and cached.

The reported ImageNet-100 experiment uses the corresponding latent response extractor with deterministic per-example Gaussian source realizations and shared response probes. Its quantization backend is adapted to the larger feature space, as described below.

Real-sample quantization. The optimization backend is adapted to dataset scale while preserving the joint covering objective in Eq. (15). For CIFAR, we use a monotone medoid-style solver that alternates sample assignment with cluster updates and projects each cluster mean back to an actual training example. Updates that increase the quantization objective are rejected. For ImageNet-100, we use a scalable minibatch centroid approximation followed by projection to the nearest real training example. Thus both implementations return valid condition–target pairs while optimizing the same joint coverage principle.

After selection, the proxy and cached response features are discarded. In the primary State-Jac protocol, the final model is trained uniformly on the selected real examples with unit loss weights.

State-Jac selection procedure. Algorithm 1 summarizes the practical selector. The zero-order feature ϕ_i is the training-space target, while ρ_i is the compressed finite-response signature. Their relative scale is calibrated only from training-set geometry.

For CIFAR, REALSAMPLEQUANTIZE denotes the medoid-style solver above; for ImageNet-100 it denotes the scalable centroid approximation followed by nearest-real-example projection. In either case, the final subset contains only real training pairs.

Randomness and uncertainty. We distinguish randomness in *subset construction* from randomness in *final-model optimization*. For the primary CIFAR-10 State-Jac result, a fixed selected manifest is reused across independent final-training seeds. The Random baseline uses independently sampled manifests, so its reported variation includes both subset and optimization variability. Entries without $\pm$ in the main tables are single-run results unless stated otherwise.

ImageNet-100 results use five independent final-training seeds. Random20 and State-Jac20 use matched final-training seeds so that comparisons share corresponding optimization randomness.

Trajectory diagnostics. These diagnostics compare a subset-trained field v_B with the corresponding full-data reference field v_F. For an evaluation pair (x, c) and source realization z, let $q_F(t)$ and $q_B(t)$ denote their rollouts under the same condition c and initial state $q_F(0) = q_B(0) = z$. This coupling controls conditioning and initialization when comparing the learned fields and their resulting trajectories. We use the per-coordinate mean squared discrepancy

$$\ell(a, b) = \frac{1}{d}\|a - b\|_2^2,$$

where d is the dimension of the compared state or velocity vector. Expectations are estimated by averages over evaluation examples and source realizations. Each time-dependent diagnostic is additionally averaged over its evaluation time grid, held fixed across the compared methods.

Own-path error and terminal drift. *Own-path error* (**Own**) measures velocity-field disagreement on states visited by the subset model. *Terminal drift* (**Drift**) measures the separation between the final states of the two matched rollouts:

$$E_{\text{own}} = \mathbb{E}_{(x,c),z,t}\left[\ell(v_B(q_B(t), t, c), v_F(q_B(t), t, c))\right], \tag{85}$$

$$D_{\text{term}} = \mathbb{E}_{(x,c),z}\left[\ell(q_B(1), q_F(1))\right]. \tag{86}$$

For Own, both fields receive the *same* subset-visited state $q_B(t)$, time t, and condition c. The metric does not compare $v_B(q_B(t), t, c)$ with $v_F(q_F(t), t, c)$, which would mix field differences with differences in the input state. Lower Own therefore indicates closer velocity predictions on states that the subset model actually visits during generation.

For Drift, each model follows its own rollout from the shared initial state. Lower Drift indicates closer final states under the shared condition and source realization. It does not measure discrepancy along the entire trajectory, and it is not reconstruction error relative to the observed target x.

The distinction follows the error decomposition in Eq. (4). Own measures the mean squared magnitude of the same-state field discrepancy along the subset rollout, whereas Drift measures terminal separation after errors accumulate and propagate. The bound in Eq. (5) also depends on state sensitivity of the reference field. Consequently, lower average Own alone does not guarantee lower Drift. Both diagnostics quantify agreement with the full-data reference; generation quality is assessed separately by FID and KID.

Reference-state field diagnostics. For the controlled cluster-mass intervention in Section 5.4, we additionally report reference-state diagnostics. *Teacher-state error* evaluates field disagreement on prescribed Rectified-Flow states, whereas *full-path error* evaluates the same discrepancy on states visited by the full-data rollout:

$$E_{\text{teacher}} = \mathbb{E}_{(x,c),z,t}\left[\ell(v_B(Q_t(x; z), t, c), v_F(Q_t(x; z), t, c))\right], \tag{87}$$

$$E_{\text{full}} = \mathbb{E}_{(x,c),z,t}\left[\ell(v_B(q_F(t), t, c), v_F(q_F(t), t, c))\right]. \tag{88}$$

Teacher-state, full-path, and Own all compare the two learned velocity fields at a common input state. They differ in the origin of that state: $Q_t(x; z)$ is prescribed by the training path, $q_F(t)$ is generated by the full-data model, and $q_B(t)$ is generated by the subset model. In particular, teacher-state error compares two learned fields; it does not compare a learned velocity with the prescribed target velocity $x - z$. Neither reference-state diagnostic measures a distance between trajectories. Because their state construction does not follow the subset model's own rollout, these errors can decrease while Own or Drift increases.

Worst-class field error. For class-labeled datasets, *worst-class field error* (**Worst**) is the maximum class-conditional Own:

$$E_{\text{worst}} = \max_{y \in \mathcal{Y}} \mathbb{E}_{(x,c),z,t|y}\left[\ell(v_B(q_B(t), t, c), v_F(q_B(t), t, c))\right].$$

The field discrepancy is first averaged within each evaluation class; the maximum is then taken across classes. Worst is therefore not the maximum error of an individual example. It reveals class-specific

Table 5: **Proxy response geometry relative to the full-data model on CIFAR-10.** Pair Corr. is Spearman correlation of pairwise response distances; NN-20 is overlap of the 20 nearest response neighbors. R_J^F is full-model response distortion normalized by random selection.

Response extractor	Pair Corr. ↑	NN-20 ↑	R_J^F ↓
Tiny proxy (default)	.79	.61	.76
Medium proxy	.84	.66	.73
Architecture-matched proxy	.87	.69	.71
Full model / oracle	1.00	1.00	.68

Table 6: **Effect of response-measurement states.** Own-geometry Corr. is rank agreement with response geometry evaluated on model-visited states. Jaccard is measured against the default reference-state subset.

Measurement states	Own-geometry Corr. ↑	Jaccard to default ↑
Reference path (default)	.72	1.00
Full-model rollout	.81	.57
Proxy rollout	.76	.61
Mixed reference/rollout	.79	.64

disagreement that may be obscured by the overall average Own. We report it for CIFAR-100 and CIFAR-10-LT. Lower Worst indicates a smaller maximum of the class-wise mean field errors; the class attaining that maximum can differ between methods.

Metric values are multiplied by 10^3 only where stated in the corresponding table captions. This scaling is for presentation and does not change the definitions.

B.2 Does the Proxy Preserve Full-Model Response Geometry?

We first test the most direct bridge between the ideal response geometry in Section 3.3 and the practical selector. Using the same sample pairs, time points, deterministic source realizations, and directional probes, we extract finite-response signatures from a proxy and from the corresponding full-data model:

$$\rho_i^p, \qquad \rho_i^F. \tag{89}$$

For sampled pair (i, j), define

$$d_p(i,j) = \left\|\rho_i^p - \rho_j^p\right\|_2^2, \qquad d_F(i,j) = \left\|\rho_i^F - \rho_j^F\right\|_2^2. \tag{90}$$

We evaluate Spearman rank agreement of pairwise distances together with local-neighborhood overlap.

Even the default tiny proxy preserves substantial full-model response ordering. Across the evaluated extractors, pairwise and neighborhood agreement increase from .79/.61 for the default proxy to .87/.69 for the architecture-matched proxy, while full-model covering distortion decreases more modestly, from .76 to .71 (oracle: .68). This supports the intended role of the proxy as a *relative geometry estimator*, rather than a surrogate generative model.

B.3 Where Should Response Geometry Be Measured?

The closed-loop analysis in Section 3.2 concerns states visited during generation, whereas the default selector evaluates responses on prescribed Rectified-Flow states. We therefore compare reference states, full-model rollout states, proxy-rollout states, and a mixed reference–rollout construction.

For each alternative we measure rank agreement with response geometry on model-visited states and subset overlap with the default State-Jac manifest. Here *Own-geometry Corr.* compares pairwise response geometry between samples; it is distinct from the Own field-error diagnostic, which compares two learned fields at the same subset-visited state.

Table 7: **Selector-seed stability on CIFAR-10.** Distortions are normalized by random selection.

Selector seed	Jaccard ↑	$R_x \downarrow$	$R_J \downarrow$
Default	1.00	.75	.72
Seed 2	.62	.76	.73
Seed 3	.60	.75	.71

Table 8: **Sensitivity of the State-Jac response sketch.** Pair Corr. is relative to the default response representation. Distortions are normalized by random selection.

Factor	Setting	Pair Corr. ↑	Jaccard ↑	$R_x \downarrow$	$R_J \downarrow$
Directions M	1	.91	.56	.76	.74
	2 (default)	1.00	1.00	.75	.72
	4	.97	.71	.75	.71
Time points T	2	.94	.61	.76	.73
	4 (default)	1.00	1.00	.75	.72
	8	.98	.73	.75	.71
ε	0.01	.96	.68	.75	.72
	0.02 (default)	1.00	1.00	.75	.72
	0.04	.94	.64	.76	.73

Moving measurement states toward actual rollouts increases agreement with own-state geometry, from .72 for the reference path to .81 for the full-model rollout and .79 for the mixed construction. At the same time, subset overlap with the default falls to .57–.64, showing that the measurement-state choice changes which examples are selected.

These diagnostics expose the relevant trade-off: rollout states are more directly aligned with model-visited geometry but are also model-dependent and require additional trajectory construction. We therefore retain prescribed states as a reproducible and inexpensive default that already exhibits substantial agreement with own-state response geometry.

B.4 Selector and Response-Sketch Stability

State-Jac contains stochasticity through its response probes and approximate quantization solver. We first fix the proxy checkpoint and regenerate subsets using different direction and solver seeds. For each manifest we measure index-level overlap and the two covering distortions used in Figure 3.

The exact selected indices differ across seeds, but their covering distortions are nearly unchanged. Low index-level overlap therefore does not imply geometric instability: multiple representative sets can cover essentially the same regions of the joint geometry.

We next vary the number of directions M, time points T, and finite-response scale ε one factor at a time around the default $(M, T, \varepsilon) = (2, 4, 0.02)$.

Across the tested range, pairwise geometry and covering distortions remain stable. One direction already preserves most pairwise ordering relative to the default sketch, and the tested direction counts yield similar covering distortions. Increasing temporal resolution beyond four points has a larger effect on the selected indices than on the achieved covering distortion. Likewise, changing ε from 0.01 to 0.04 leaves the induced geometry nearly unchanged. This is qualitatively consistent with Proposition 5: although the one-sided response has first-order pointwise derivative error, central symmetry removes the leading correction in the expected squared *pairwise response geometry*.

These results support the deliberately small default sketch: State-Jac recovers stable relative response geometry without turning selection into a second expensive training procedure.

Table 9: **Cluster-mass control on CIFAR-10.** The selected representatives are fixed; only the empirical sampling distribution changes. Values are reported from the matched single-run comparison used for the intervention, whereas State-Jac entries in Tables 1–2 report means and standard deviations across independent final-training runs.

Geometry	Sampling	FID ↓	Teacher ↓	Full-path ↓	Own ↓	Drift ↓
Endpoint	Uniform	15.532	.02991	.02777	.02120	.01201
Endpoint	Cluster mass	18.238	.02333	.02062	.02838	.01757
State-Jac	Uniform	15.102	.02616	.02687	.02075	.01175
State-Jac	Cluster mass	19.426	.02329	.02023	.02861	.01776

B.5 Cluster-Mass Sampling Exposes the Closed-Loop Gap

Figure 4(a–b) summarizes a controlled intervention that changes the empirical training measure while keeping the selected examples fixed. For the CIFAR medoid set $\mathcal{B}$, let $\mathcal{C}_b$ denote the assignment cell represented by b. The primary protocol samples the K selected representatives uniformly, whereas the intervention uses

$$p_b = \frac{|\mathcal{C}_b|}{N}. \tag{91}$$

Thus the selected examples and per-example CFM loss are unchanged; only the empirical training measure differs.

For State-Jac, cluster-mass sampling lowers teacher-state and full-path errors by 11.0% and 24.7%, respectively, yet increases Own by 37.9% and Drift by 51.1%; FID degrades from 15.102 to 19.426. These percentage changes are computed relative to the matched uniform row in Table 9, rather than the multi-run averages in the main results tables. Endpoint representatives exhibit the same qualitative reversal.

This provides a controlled empirical example in which improved reference-state agreement coexists with worse Own, Drift, and FID. Changing the training measure can improve agreement on prescribed or full-model-generated states while increasing disagreement on states visited by the subset model and increasing terminal separation. The result is consistent with Eq. (4) and motivates retaining uniform sampling in the primary method.

B.6 Geometry–Outcome Alignment

Finally, we test whether the two covering geometries align with the distinct downstream behaviors motivating State-Jac. For subset $\mathcal{B}$,

$$R_x(\mathcal{B}) = \frac{1}{N}\sum_{i=1}^{N} \min_{b\in\mathcal{B}} d_x^2(i,b), \qquad R_J(\mathcal{B}) = \frac{1}{N}\sum_{i=1}^{N} \min_{b\in\mathcal{B}} d_J^2(i,b). \tag{92}$$

Here d_J is the practical distance induced by the standardized finite-response signature used for selection, rather than an explicitly materialized Jacobian distance. The two minima are taken separately, so the nearest representative can differ between endpoint and response geometry. Equation (92) defines the raw distortions; reported normalized values divide each distortion by its corresponding Random reference value at the same subset budget.

The statistical unit is a *unique subset manifest*, rather than an individual sample or final-training seed. When multiple final-training runs reuse the same manifest, we first average their downstream metrics. We restrict the analysis to uniform final-training methods, excluding weighted and cluster-mass variants so that changes in training measure do not confound selection geometry.

The named selectors illustrate the qualitative division observed in Table 2: endpoint- and residual-based selectors attain relatively low zero-order distortion, Response-only provides the strongest pure response coverage, and State-Jac lies in the favorable region of both.

The dotted lines in Figure 3(c–d) are visual guides; quantitative conclusions use manifest-level rank statistics over the full analysis pool rather than only the named methods in Table 10. After deduplicating manifests and averaging repeated final-training runs, the analysis contains $n = 30$

Table 10: **Normalized CIFAR-10 covering distortions for representative selectors shown in Figure 3(c–d).** Random is normalized to one separately for each geometry.

Method	$R_x \downarrow$	$R_J \downarrow$
Random	1.00	1.00
DINOv2 K-Medoids	1.16	1.10
DINOv2 K-Center	1.29	1.17
EL2N-FM High	1.08	1.19
Moderate-DS	.90	.95
Endpoint	.78	.93
Response-only	1.04	**.69**
Residual	.82	.86
State-Jac	**.75**	.72

Table 11: **Manifest-level geometry–outcome alignment on CIFAR-10.** Entries are Spearman rank correlations over $n = 30$ unique uniform subset manifests.

Geometry	Downstream outcome	Spearman $\rho \uparrow$
Endpoint R_x	FID	**.89**
Response R_J	Terminal drift	**.93**

unique uniform subset manifests. We focus on the two geometry–outcome relationships motivated by the zero-/first-order decomposition and visualized in Figure 3.

Both relationships show strong positive rank association: endpoint covering distortion closely tracks FID ($\rho = .89$), while response covering distortion closely tracks terminal drift ($\rho = .93$). These associations are consistent with the complementary roles suggested by the component ablations: endpoint coverage concerns represented data support, while response coverage concerns local state sensitivity.

These relationships are empirical associations rather than deterministic performance guarantees; final behavior also depends on optimization and on the state distribution induced by subset retraining. Their role is to test whether the two components of the selection geometry are empirically associated with the distributional and dynamical behaviors that motivated their construction.

B.7 Takeaway

These experiments directly audit the practical gap between the ideal geometry of Section 3 and the one-shot selector of Section 4. A small proxy preserves much of the full model's relative response ordering; prescribed states provide a stable alternative to more model-dependent rollout measurements; the compact finite-response sketch is robust to moderate changes in probes, time resolution, perturbation scale, and selector seed; and its zero- and first-order covering distortions align with distinct downstream outcomes.

These results support the practical premise that State-Jac need not reconstruct the full rollout operator or the full state Jacobian, provided that the resulting relative zero-order supervision and first-order local-response geometry remain informative for selection before the expensive final training run.