RECONSTRUCTION IS NOT TRANSPORTABILITY: TRANSITION-STATE-AWARE TOKENIZATION FOR REACTION PATHWAY FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Mapping the pathway that carries reactants through the transition state to products is a bottleneck in mechanistic analysis and reaction screening, and the fastest learned methods now produce a complete pathway in one pass from the endpoint geometries alone. Generating it in a learned latent space separates geometric encoding from transport, but a latent trained only to reconstruct is not the one a flow transports best. We call this the *representation–transport mismatch*: a pathway is recovered from its latent to 0.052 Å while that latent stays hard for the flow to predict. We propose REACFLOW, a two-stage framework that shapes the pathway latent for transport rather than for reconstruction alone. Its autoencoder is pretrained with a head that uses the latent and endpoints to predict projected-force responses inspired by climbing-image NEB, and an interval-completion objective whose extreme case is the endpoint interpolation the generator starts from. We then freeze it and train a conditional flow-matching model to carry the interpolation latent to the target latent; generation uses four velocity-field evaluations followed by the frozen decoder, with no energy or force evaluations. We test REACFLOW on two benchmarks, Transition1x and Halo8, and it attains the lowest geometry error of any method compared on both (0.215/0.175 Å TS-RMSD and 0.155/0.165 Å IRC-RMSD), improving generation by 24.6–36.7% over reconstruction-only pretraining while clean reconstruction stays flat at 0.052–0.056 Å.

1 INTRODUCTION

Predicting a reaction pathway—the continuous sequence of geometries carrying reactants through the transition state to products—underpins mechanistic analysis, barrier estimation, and reaction screening. Chain-of-states methods such as the nudged elastic band and the string method remain the reference procedure, but they reach a pathway only by evaluating energies and forces on every image, over and over, which puts them out of reach at screening scale (Henkelman et al., 2000; E et al., 2002). Most learning-based methods remove that loop by conditioning on the endpoints alone: from reactant and product geometries, MEPIN predicts deviations from a geometric interpolation, and Drift-React generates the complete pathway with an $\mathrm{SE}(3)$-equivariant drifting field (Nam et al., 2025; Schlama & Schwaller, 2026). Endpoint conditioning is what makes these methods fast, and also what makes them hard: the endpoints fix where a pathway starts and ends, but not the coordinated motion in between—the part that sets the barrier.

These methods fill that gap with an interpolation—in the simplest case a straight line, along which atoms move independently, with none of the coordination that carries the system over the barrier. A coordinate-space network must supply that coordination while carrying the interpolation onto the true pathway. Geometric pretraining splits these two jobs, as in molecular trajectory and protein generation (Iyengar et al., 2026; Ni et al., 2026): a tokenizer first learns the joint geometry across atoms and path positions, and a conditional Flow then learns latent transport there. The tokenizer's default objective is reconstruction—can a decoder recover an observed path from its latent—but generation asks whether the Flow can predict that latent from the endpoints alone. The encoder sees the saddle region when compressing a reference pathway; at generation time, the Flow must predict the target latent from the endpoints alone. Nothing in the reconstruction loss constrains the second question, so a latent can preserve the intermediate geometry and still be hard to predict: across our

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.

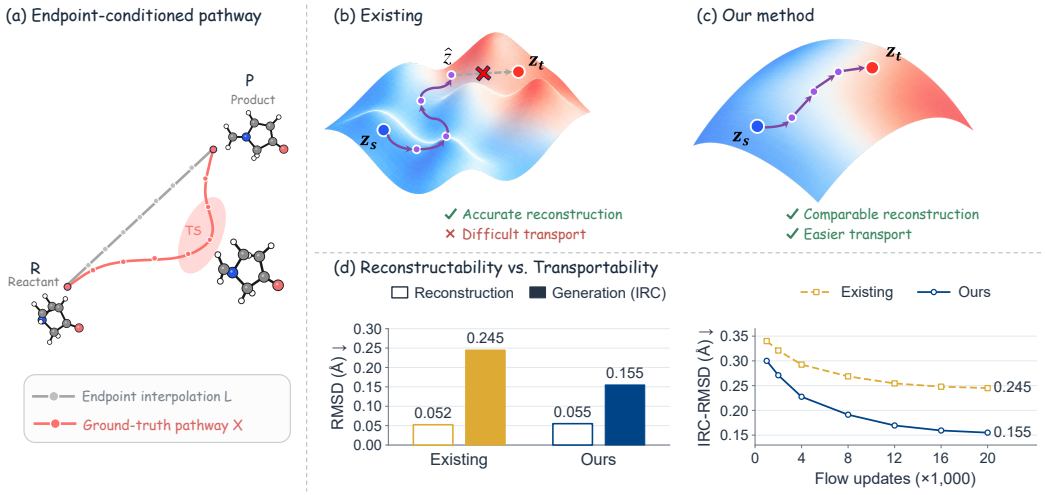

Figure 1: Representation–transport mismatch in endpoint-conditioned pathway generation. (a) Endpoint interpolation misses nonlinear intermediate motion. (b,c) Accurate reconstruction need not imply easy transport; auxiliary pretraining preserves reconstruction while improving transportability. (d) Similar reconstruction errors yield different generation errors, so reconstruction alone is insufficient for selecting a pathway tokenizer. Existing and Ours denote reconstruction-only and full pretraining on Transition1x, respectively. Reconstruction uses raw all-frame RMSD; generation uses aligned IRC-RMSD.

pretraining objectives clean reconstruction spans only 0.052–0.056 Å while endpoint-conditioned generation error falls by 24.6–36.7%. We call this a *representation–transport mismatch* (Figure 1), the reaction-pathway form of the reconstruction–generation gap in latent generative models (Yang et al., 2026).

A reaction pathway is not an arbitrary sequence of geometries; it is the output of a saddle search. We ask two things of the latent: that it keep what that search moves, and that it be predictable from the endpoints alone. For the elementary reactions here, CI-NEB makes the first concrete: at every image it relaxes the transverse force, and at the highest-energy image it reverses the tangential force so the image climbs toward the saddle (Henkelman & Jónsson, 2000; Henkelman et al., 2000). *Reaction relevance* requires these corrections to stay recoverable from the compressed latent and endpoints by a shallow readout; search-response supervision trains that readout. Endpoint-anchored interval completion implements the second, *conditional predictability*, as masked modeling (He et al., 2022; Zhao et al., 2026): we rebuild the complete pathway after a contiguous internal interval is replaced by interpolation between its retained boundaries. Replacing every internal frame makes that input exactly the endpoint-linear source, so pretraining and generation see the same input.

To overcome these two challenges, we propose REACFLOW: one tokenizer trained on both objectives beside self-reconstruction, then frozen while a conditional flow-matching model—*the Flow*—learns transport in its latent space. Its design turns on a constraint: our labels for what a search moves are built from energies and forces, which generation may not touch; REACFLOW spends that physics in pretraining, on records the search already produced. Both reach the same encoder—completion also trains the decoder—but do different work: search response targets what survives compression, completion what is recoverable from the endpoints alone. These objectives shape the encoder before it is frozen for flow matching. Freezing fixes that target: we discard the head, and the Flow trains against latents it cannot reshape; no pretraining loss directly matches source and target latents, so the Flow inherits a fixed problem, not its answer. We summarize our contributions as follows.

- **We measure the representation-transport mismatch.** At fixed Flow capacity and training budget, clean reconstruction moves only from 0.052 to 0.056 Å across pretraining objectives, while endpoint-conditioned generation error falls 24.6–36.7%.
- **Search-response supervision.** A shallow head predicts CI-inspired projected-force responses from candidate latents and endpoint conditions, encouraging reaction-relevant cor-

rections to remain recoverable after compression. The head is discarded before transport, so generation needs no energies or forces; generic force labels instead of the CI-inspired target cost $0.155 \rightarrow 0.175$ Å IRC-RMSD.

- **Endpoint-anchored interval completion.** Completing a missing interval from retained boundaries exposes the tokenizer to partial context, including the endpoint-linear source used at generation. Unstructured masking instead costs $0.155 \rightarrow 0.180$ Å.
- **Analysis and results.** We give a necessary latent-separation condition for accurate search-response prediction, and bound the error of an endpoint-conditioned latent predictor constructed by completion and re-encoding. On Transition1x the full model reduces TS-RMSD by 17.9%, IRC-RMSD by 57.4%, and Fréchet distance by 65.2% against the strongest baseline in each column, with consistent geometric gains on Halo8.

2 Related Work

Electronic-Structure Pathway Search. Electronic-structure approaches include IRC path following from a transition state (Fukui, 1981) and chain-of-states optimization, such as NEB, climbing-image NEB, and string methods (Jónsson et al., 1998; Henkelman et al., 2000; E et al., 2002; Peters et al., 2004). These procedures require repeated energy and force evaluations. Machine-learned potentials reduce their cost, as in NeuralNEB (Schreiner et al., 2022b), while implicit neural representations optimize paths with path-projected energy gradients (Ramakrishnan et al., 2025); both still iterate at solve time. Our search-response supervision is confined to pretraining; generation evaluates no energies or forces.

Endpoint-Conditioned Pathway Generation. Endpoint-conditioned generators differ in what they transport and how that space is chosen. MEPIN and Drift-React map endpoint interpolations to full pathways in Cartesian coordinates (Nam et al., 2025; Schlama & Schwaller, 2026). For transition-state generation, OA-ReactDiff conditions on 3D endpoint geometries (Duan et al., 2023), whereas TSDiff conditions on 2D reaction graphs (Kim et al., 2024). React-OT uses an endpoint-informed source for deterministic transport, and MolGEN uses conditional flow matching (Duan et al., 2025; Tuo et al., 2025). TS-DFM transports a fixed distance-geometry representation rather than Cartesian coordinates (Luo et al., 2025). We instead learn a pathway representation before transport; the contribution is how it is shaped, with comparable reconstruction but 25–37% differences in generation error.

Representation Quality in Latent Generative Models. In two-stage latent generation, reconstruction alone is insufficient for selecting a tokenizer: a latent that reconstructs accurately need not be easy for a denoiser or flow to predict (Yang et al., 2026). Predictive and masked objectives emphasize inference from partial context (He et al., 2022; Zhao et al., 2026), including in molecular-trajectory and protein pretraining (Iyengar et al., 2026; Ni et al., 2026). For reaction pathways, the missing content includes the saddle-region geometry absent from endpoint interpolation. We use reconstruction-only pretraining as a controlled baseline and test whether search-response supervision and interval completion improve transport at comparable reconstruction accuracy.

3 Method

ReacFlow has two stages. We first pretrain a shared pathway autoencoder—the *tokenizer*—with reconstruction, search-response supervision, and interval completion, then freeze it and train an independent endpoint-conditioned flow-matching model in its latent space (Figure 2). Appendices B and D give architecture and optimization details.

Problem setup. Both Transition1x and Halo8 use paths $X \in \mathbb{R}^{T\times N\times 3}$ with fixed $T = 10$: two atom-mapped endpoints $R = X_0$ (reactant) and $P = X_9$ (product), plus eight internal frames $X_1, \ldots, X_8$. Here N counts valid atoms; batches are padded to a common width, and all losses and statistics exclude padding. Frames share a reactant-centered coordinate system. The endpoint-linear source is

$$L_k = (1 - s_k)R + s_k P, \qquad s_k = k/(T-1). \tag{1}$$

where $s_k \in [0, 1]$ is the normalized position of frame k along the path. The datasets also retain the intermediate paths visited by the reference search, each with its energies and forces; we write

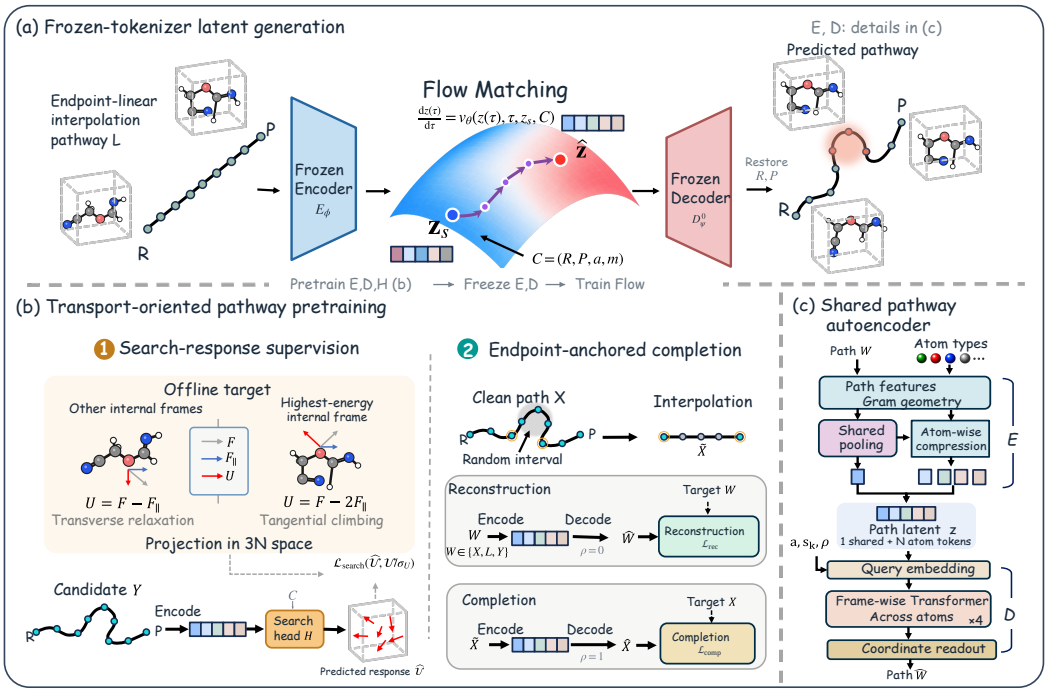

Figure 2: Transport-oriented pretraining and generation. (a) At generation, the frozen encoder maps the endpoint-linear source L to a latent, the conditional flow model transports it toward the target latent, and the frozen decoder returns coordinates; the known endpoints are then overwritten. (b) That latent space comes from pretraining, where search-response supervision and endpoint-anchored interval completion accompany self-reconstruction. (c) Both stages share one autoencoder: trajectory Gram features compress a pathway into one shared token plus one token per atom.

Y for such a candidate path, and reusing it costs no new electronic-structure evaluations. We write $W \in \{X, L, Y\}$ for a self-reconstruction input and $C = (R, P, a, m)$ for the condition available to the search head and Flow, collecting the endpoints with atom types a and the padding mask m.

3.1 The Full-Pathway Tokenizer

We encode an entire pathway into a single latent rather than one latent per frame, so the Flow learns one source-to-target transport instead of T coupled ones to enforce the temporal consistency. **Encoder and latent structure.** The encoder E_ϕ reads full coordinates. Each atom's trajectory $Q_i \in \mathbb{R}^{T\times 3}$ enters through the Gram matrix $Q_i Q_i^\top$, which is invariant to global rotation; an MLP turns it into scalar coefficients that recombine the original coordinate vectors, so the latent is equivariant by construction. A mean-pooled global vector is appended to every atom's input, so each atom-wise latent sees the whole molecule. With $K = 4$ vector channels,

$$z_W = (G, b, \{Z_i\}_{i=1}^N), \qquad G, Z_i \in \mathbb{R}^{K\times 3}, \quad b \in \mathbb{R}^K. \tag{2}$$

One global token carries the pooled vector G and scalar b; one token per atom carries Z_i, the sum of a fixed projection of atom i's trajectory onto K orthogonal temporal modes and a learned, zero-initialized geometric correction.

Shared frame-query decoder. The decoder D_ψ^ρ reads the latent, atom metadata, and the frame position s_k. Its flag selects self-reconstruction/generation ($\rho = 0$) or completion ($\rho = 1$) and enters only the decoder, so both modes share one latent. Four Transformer layers attend across atoms within each frame; weights are shared across frames, not layers. No raw coordinates or interpolation path reach the decoder. For $O \in \mathrm{O}(3)$ and an atom permutation Π, define $(\mathcal{G}_{O,\Pi}X)_k = \Pi X_k O^\top$, with atom types and masks permuted accordingly. The noiseless tokenizer satisfies $D_\psi^\rho(E_\phi(\mathcal{G}_{O,\Pi}X)) = \mathcal{G}_{O,\Pi} D_\psi^\rho(E_\phi(X))$. Translation is handled by reactant centering and

subsequent restoration. Isotropic latent noise preserves equivariance in distribution, not samplewise equality for independently drawn perturbations. The decoder receives neither R nor P and reconstructs all T frames from the latent; only generation overwrites the endpoints with the true R and P. Equation 11 in Appendix B specifies the synthesis.

3.2 Search-Response Supervision

A pathway is not just a shape to reconstruct: it is the answer to a search problem. We therefore attach a deliberately shallow head to the latent and require it to predict a CI-inspired projected-force response to a candidate path. The shallow head encourages search-response information to remain accessible from the latent, conditional on the endpoints. Let Y be a recorded candidate path with its recorded energies and forces. The energies define an energy-weighted unit tangent t_k in the full $3N$-dimensional configuration space (Henkelman & Jónsson, 2000). Write F_k for the recorded force, $F_k^{\parallel} = \langle F_k, t_k \rangle_{3N} t_k$ for its tangential part, and $k^\star$ for the highest-energy internal frame. Following the climbing-image construction (Henkelman et al., 2000), the target *response* is

$$\mathcal{U}_k(Y) = \begin{cases} F_k - 2F_k^{\parallel}, & k = k^\star, \\ F_k - F_k^{\parallel}, & 1 \le k \le T-2,\ k \ne k^\star, \\ 0, & k \in \{0, T-1\}. \end{cases} \tag{3}$$

Non-climbing frames keep only the transverse force; the frame $k^\star$ reverses its tangential component. We omit the NEB spring term, so $\mathcal{U}$ is a CI-inspired response rather than a full optimizer update. The labels are read off existing training records and cost no new energy or force evaluations (Appendix A).

The head $H_\omega(E_\phi(Y), C)$ predicts the normalized response $\mathcal{U}(Y)/\sigma_U$, where σ_U is a fixed training-set response RMS. It combines latent and endpoint vectors through invariant coefficients and never sees uncompressed internal coordinates; energies, forces, tangents and the peak index enter only as labels. Theorem 2 (Appendix E) makes this precise: for identical endpoint conditions, a response difference exceeding the two readout errors requires latent separation, with its scale set by head sensitivity. This concerns retained response information, not transportability.

3.3 Endpoint-Anchored Interval Completion

At generation time the model sees only the endpoint interpolation L, so we train the encoder to rebuild a full pathway from partial, endpoint-anchored context (He et al., 2022; Zhao et al., 2026). We replace a contiguous span $\mathcal{M}$ of internal frames by interpolation between the two frames that bound it. Sampling has two stages: a span length, uniform over $1, \ldots, T-2$, then a uniform admissible start. Endpoints are always retained, and sampling uses no energies, forces, or peak index. With u, v the nearest retained boundary indices, the corrupted path is

$$\widetilde{X}_k = \begin{cases} \dfrac{v-k}{v-u} X_u + \dfrac{k-u}{v-u} X_v, & k \in \mathcal{M}, \\ X_k, & k \notin \mathcal{M}. \end{cases} \tag{4}$$

Pretraining contains the generation-time input. Setting $u = 0$ and $v = T-1$ in Equation 4 collapses it to Equation 1: masking every internal frame reproduces the endpoint-linear source L exactly. That span is drawn with probability $1/(T-2)$, so pretraining explicitly covers the endpoint-only input used at generation, without physical labels. The completion decoder and the Flow nevertheless learn different maps. Completion predicts $\widehat{X}^{\,c} = D_\psi^1(E_\phi(\widetilde{X}))$. A single latent $E_\phi(L)$ must therefore decode two ways: to L itself when $\rho = 0$, and to the target pathway X when $\rho = 1$, with the flag as the only difference. Neither the flag nor the mask reaches the encoder, and no pretraining loss compares $E_\phi(L)$ with $E_\phi(X)$, so completion does not hand the Flow its answer.

Joint objective. Let $\ell_{\mathcal{A}}$ be coordinate MSE over valid atoms and frames $\mathcal{A}$: all frames (all), internal frames (int), the masked span ($\mathcal{M}$), or its retained complement ($\mathcal{V}$). We jointly optimize

$$\begin{aligned}
\mathcal{L}_{\text{rec}} &= \tfrac{1}{3} \sum_{W \in \{X,L,Y\}} \mathbb{E}\, \ell_{\text{all}}\big(D^0_\psi(E_\phi(W)+\epsilon), W\big), \\
\mathcal{L}_{\text{search}} &= \mathbb{E}\, \ell_{\text{int}}\big(H_\omega(E_\phi(Y), C), \mathcal{U}(Y)/\sigma_U\big), \\
\mathcal{L}_{\text{comp}} &= \tfrac{1}{2}\mathbb{E}\big[\ell_{\mathcal{M}}(\widehat{X}^{\,c}, X) + \ell_{\mathcal{V}}(\widehat{X}^{\,c}, X)\big], \\
\mathcal{L}_{\text{tok}} &= \mathcal{L}_{\text{rec}} + \lambda_s \mathcal{L}_{\text{search}} + \lambda_c \mathcal{L}_{\text{comp}}.
\end{aligned} \tag{5}$$

Reconstruction is the only term that perturbs the latent, by isotropic noise ϵ with standard deviation 0.01. Averaging masked and retained frames separately prevents a long masked span from dominating $\mathcal{L}_{\text{comp}}$ through frame count alone. Both auxiliaries train the encoder; completion also trains the decoder. The completion loss bounds endpoint-only risk up to a coverage factor; encoder stability transfers this bound to latent prediction. Below, $\|\cdot\|_z$ is Euclidean over valid latent entries; Appendix E supplies stability bounds and proofs.

Theorem 1 (Completion-induced latent predictability)**.** *Fix the tokenizer and a reaction distribution with finite $\mathcal{L}_{\text{comp}}$. Let the conditional all-internal-mask probability satisfy $q(X) \geq p_{\min} > 0$; our sampler gives $q(X) = p_{\min} = 1/(T-2)$. Define $Q(C) = P_C D^1_\psi(E_\phi(L(C)))$, where P_C restores the known endpoints. If E_ϕ has a common finite L_E-Lipschitz bound from all-frame per-coordinate RMS to the latent norm on neighborhoods of $[Q(C), X]$, the endpoint-conditioned predictor $g(C) = E_\phi(Q(C))$ satisfies*

$$\mathbb{E}_X \|g(C) - E_\phi(X)\|_z^2 \leq \frac{2L_E^2(T-2)}{T p_{\min}} \mathcal{L}_{\text{comp}} =: \varepsilon_c^2. \tag{6}$$

3.4 Endpoint-Conditioned Latent Transport

Stage two learns the map the tokenizer was built for: from the latent of the endpoint-linear path to the latent of the true pathway. We freeze the tokenizer, discard the search head H_ω, re-encode every training pathway with the frozen encoder, and train the flow model and its endpoint network from scratch. Transport time τ is distinct from path position s_k:

$$z_s = E_\phi(L), \qquad z_t = E_\phi(X), \qquad z_\tau = (1-\tau)z_s + \tau z_t, \qquad u^\star = z_t - z_s. \tag{7}$$

The source z_s encodes the endpoint-linear input covered by all-internal-frame completion. We parameterize the conditional velocity field $v_\theta(z_\tau, \tau, z_s, C)$ with six directional message-passing blocks over the shared token and atom-wise tokens, exchanging scalar and vector information at every layer. Velocity errors are divided by fixed per-channel RMS scales S, measured once on clean training latents $E_\phi(X)$ and never updated. With $\tau \sim \mathcal{U}(0,1)$, the flow-matching objective (Lipman et al., 2023; Liu et al., 2023) is

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\tau,X} \left\| \big(v_\theta(z_\tau, \tau, z_s, C) - u^\star\big)/S \right\|_{\text{br}}^2. \tag{8}$$

The branch norm weights the shared and atom-wise contributions equally, keeps scalar and vector channels separate, and excludes structural zeros (Appendix B).

Flow training has two objective phases. The first $4{,}096$ updates optimize Equation 8 alone (***FM only***). We then retain flow matching and add decoded-geometry supervision (***FM + geometry***): differentiable Heun rollouts are decoded through frozen D^0_ψ, with endpoints restored by P_C, and supervised using both rigidly aligned and raw coordinate errors. During the final $4{,}096$ geometry-supervised updates, we reverse reaction direction with probability 0.5 at a reduced learning rate (Appendix D); inference remains unchanged. The tokenizer remains frozen throughout; these gradients update the Flow, not the encoder or decoder.

Inference. Generation uses a deterministic rollout from the linear-path latent $E_\phi(L)$ with two Heun steps (NFE $= 4$). We decode with $\rho = 0$, undo reactant-centering, and restore the known endpoints (P_C).

Why good reconstruction is not enough. Let $\widehat{z}(C)$ be the actual numerical Flow output from endpoints C. Under its stability and finite-risk assumptions, Theorem 3 (Appendix E) gives the upper bound

$$\sqrt{\mathcal{R}_{\text{gen}}} \leq \sqrt{\mathcal{R}_{\text{rec,clean}}} + L_D(\varepsilon_c + \delta_F), \qquad \delta_F = \big(\mathbb{E}\|\widehat{z}(C) - g(C)\|_z^2\big)^{1/2}, \tag{9}$$

Table 1: Full-pathway generation (mean ± SD; lower is better). **Bold**/underline: best/second-best. All baseline results, including Drift-React, are reported by Schlama & Schwaller (2026); REACFLOW is evaluated using their released code under the same protocol. *NeuralNEB retrained in that work. Protocols: Appendix C.1.

		Geometry (Å) ↓						**Energy MAE** (eV) ↓			
Dataset	Method	TS-RMSD		IRC-RMSD		Fréchet		$\lvert\Delta E^{\ddagger}\rvert$		E_{image}	
Transition1x	Linear	0.432	±0.196	0.364	±0.115	0.748	±0.230	1.881	±1.203	0.917	±0.279
	IDPP	0.432	±0.196	0.364	±0.115	0.748	±0.230	1.896	±1.205	**0.915**	±0.278
	Geodesic	0.432	±0.193	0.364	±0.111	0.747	±0.230	1.910	±1.210	0.916	±0.278
	NeuralNEB*	0.682	±0.322	0.536	±0.203	0.789	±0.248	1.977	±2.259	1.319	±1.617
	Drift-React	0.262	±0.161	0.386	±0.213	0.797	±0.447	1.631	±1.380	3.214	±2.069
	REACFLOW	**0.215**	±0.143	**0.155**	±0.118	**0.260**	±0.215	**1.503**	±1.267	1.047	±0.934
Halo8	Linear	0.334	±0.190	0.205	±0.108	0.349	±0.201	8.826	±51.490	3.937	±16.814
	IDPP	0.337	±0.184	0.207	±0.108	0.351	±0.196	1.651	±1.246	1.306	±0.701
	Geodesic	0.316	±0.206	0.205	±0.116	0.338	±0.202	1.247	±1.014	1.055	±0.558
	NeuralNEB*	0.761	±0.555	0.477	±0.324	0.793	±0.595	8.031	±32.313	4.638	±17.575
	Drift-React	0.185	±0.149	0.213	±0.189	0.367	±0.327	0.914	±2.980	0.621	±1.145
	REACFLOW	**0.175**	±0.128	**0.165**	±0.126	**0.280**	±0.231	**0.802**	±1.164	**0.556**	±0.887

where $\mathcal{R}_{\text{gen}}$ and $\mathcal{R}_{\text{rec,clean}}$ are expected all-frame per-coordinate MSEs for endpoint-restored generation and noiseless target reconstruction. Here ε_c bounds g's latent prediction error, while δ_F measures the Flow output's discrepancy from g; both are scaled by decoder sensitivity L_D. Unlike the clean risk, training loss $\mathcal{L}_{\text{rec}}$ mixes $X/L/Y$ inputs and adds latent noise. The theorem neither bounds δ_F by $\mathcal{L}_{\text{FM}}$ nor proves a nonzero reconstruction–generation gap: it gives an upper bound, not an additive decomposition. Completion and re-encoding construct g only for analysis, not generation. We therefore test the mismatch empirically through separate reconstruction, latent-error, and decoded-geometry measurements.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data and controls. We evaluate on Transition1x (Schreiner et al., 2022a) and Halo8 (Lee et al., 2025), using ten-frame pathways (two endpoints and eight internal frames) with fixed atom correspondence. Transition1x uses 9,000 training and 1,073 evaluation reactions. Halo8 adds halogen-containing chemistry and follows the reactant-grouped split procedure released with Drift-React; we train separately on each dataset (Appendix A). All baseline results in Table 1, including Drift-React, are taken from Schlama & Schwaller (2026). On both datasets, we evaluate REACFLOW using their released evaluation code under the same benchmark protocol (Appendix C.1). Controlled tokenizer variants share $X/L/Y$ coordinates, atom mappings, partitions, architecture, and optimization budgets: 8,192 tokenizer updates followed by 20,480 updates of a freshly initialized Flow; physical labels are available only to the tested objective.

Metrics. We follow the released Drift-React evaluation code: TS-RMSD compares aligned generated and reference structures at the same reference-energy maximum; IRC-RMSD averages aligned errors over all ten frames, including endpoints; discrete Fréchet compares complete paths under order-preserving matching. Tables 2 and 3 use the same generation-geometry protocol as Table 1 (Appendix C.1). We also report reference-peak barrier error $|\Delta E^{\ddagger}|$, per-image energy MAE, and clean-path reconstruction RMSD. Reported geometry metrics use atom-count (N) normalization; Appendix C distinguishes benchmark metrics, reconstruction, and case diagnostics from per-coordinate training losses and theoretical risks.

4.2 PATHWAY GENERATION QUALITY

REACFLOW attains the lowest mean error on all three geometry metrics on both datasets (Table 1), and its margin grows with the length scale of the comparison: on Transition1x it improves on the

Table 2: Reconstruction fidelity does not predict transportability. Transition1x (N-normalized, Å, ↓); best **bold**.

	Recon.	Generation		
Pretraining objective	Clean	TS-RMSD	IRC-RMSD	Fréchet
Reconstruction (R)	**0.052**	0.285	0.245	0.410
+ search response (S)	0.054	0.235	0.205	0.340
+ interval compl. (C)	0.056	0.225	0.190	0.310
+ both (R+S+C)	0.055	**0.215**	**0.155**	**0.260**

Table 3: Each control alters one component of R+S+C. Transition1x (Å, ↓); best **bold**.

	Generation		
Control	TS-RMSD	IRC-RMSD	Fréchet
S → physical force	0.245	0.175	0.300
C → random mask	0.250	0.180	0.315
C → decoder only	0.245	0.175	0.305
Full (R+S+C)	**0.215**	**0.155**	**0.260**

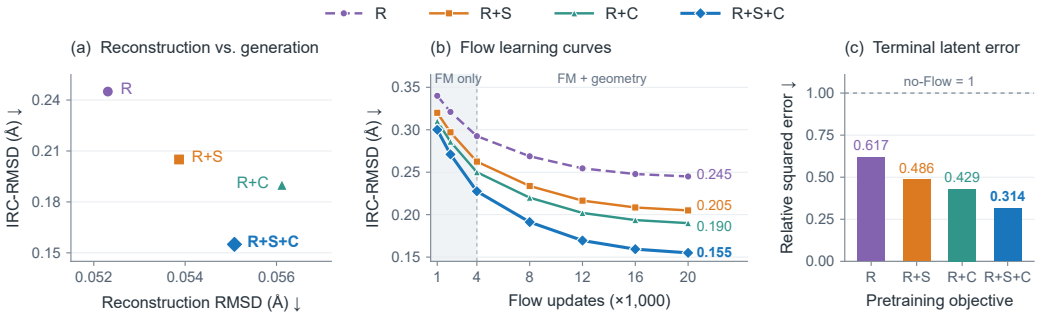

Figure 3: Similar reconstruction fidelity, different transport performance (Transition1x). (a) Clean reconstruction versus generated IRC-RMSD. (b) FM-only warm-up (shaded), followed by FM plus decoded-geometry supervision; the tokenizer remains frozen. (c) Terminal latent squared error relative to each tokenizer's no-Flow baseline (Equation 16). Geometry is N-normalized; (c) is dimensionless. Lower is better.

best baseline by 17.9% at the transition state, by 57.4% averaged along the pathway, and by 65.2% on whole-path shape. The largest gains are in whole-path geometry, while Transition1x barrier MAE falls by 7.8%. IDPP achieves the lowest per-image energy MAE on Transition1x. Margins on Halo8 are tighter (5.4%, 19.5% and 17.2% on the three geometry metrics, 12.3% and 10.5% on the two energy errors), with linear interpolation already a stronger baseline there than on Transition1x.

4.3 What Makes a Latent Transportable?

We conduct ablation studies on four tokenizer variants, which differ only in pretraining objective: reconstruction alone (R), plus search-response supervision (R+S), plus endpoint-anchored interval completion (R+C), and both (R+S+C), the objective ReacFlow uses.

Reconstruction does not rank generation quality. R reconstructs best yet generates the least accurate paths (Table 2, Figure 3a). Clean reconstruction remains at 0.052–0.056 Å across objectives, while R+S+C improves the three generation metrics by 24.6–36.7% over R.

Transport gains persist across the tested Flow budget. Under the same training schedule, R+S+C has the lowest IRC-RMSD at every displayed checkpoint, during both FM-only and FM-plus-geometry training (Figure 3b). At the final displayed budget, it reaches 0.155 Å versus R's 0.245 Å.

Smaller transport residuals before decoding. R+S+C retains 31.4% of its own no-Flow squared latent discrepancy, versus 61.7% for R (Figure 3c). These tokenizer-relative scores reproduce the generation ordering, but are not absolute latent distances in a common space.

Each substitution gives back part of the gain. Replacing the projected response $\mathcal{U}$ with the raw force F, the contiguous interval with scattered masks, or completion's encoder gradients with decoder-only updates each gives back 22–50% of the R-to-full gain, scattered masking the most, and none recovers the full objective (Table 3).

On Transition1x, latent transport with the same frozen tokenizer improves over direct endpoint completion by 13.7%, 22.9%, and 21.9% in TS-RMSD, IRC-RMSD, and Fréchet, respectively (Appendix C.4); completion alone does not match the full generation route.

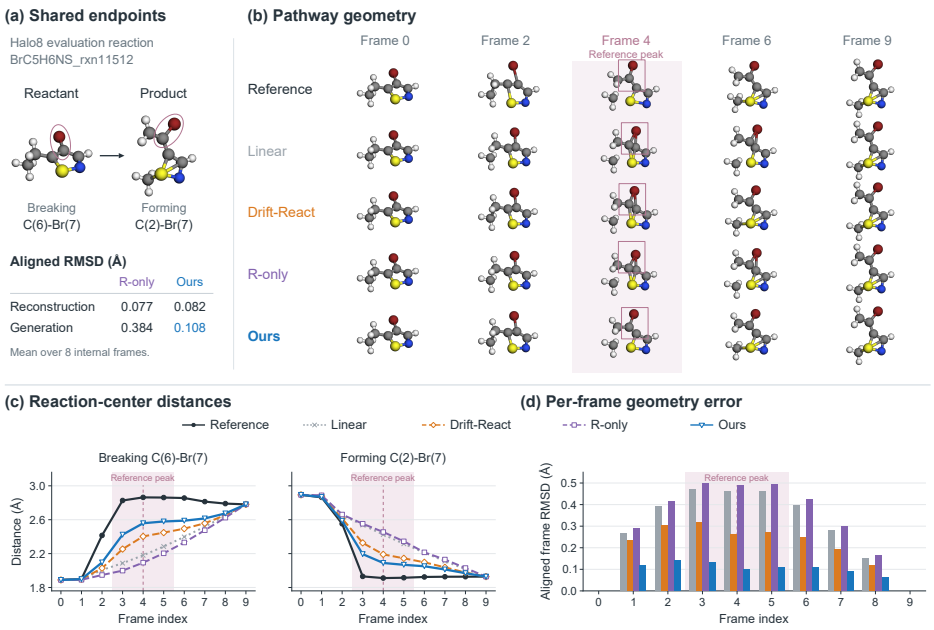

Figure 4: On the Halo8 C–Br reaction, all methods lag reference bond exchange; REACFLOW lags least. (a) Shared endpoints and clean-reference reconstruction/generation RMSDs averaged over eight internal frames. (b) Paths at frames $0, 2, 4, 6, 9$, with reaction-center boxes at reference peak frame 4; visual agreement does not certify a generated transition state. (c) C–Br distances. (d) Per-frame RMSD. RMSDs are proper-rotation-aligned and N-normalized; shading marks frames 3–5 and dashed lines frame 4.

4.4 REACTION-CENTER ANALYSIS OF A SINGLE REACTION

Aggregate scores average over reactions and over positions along a pathway, and so cannot show *where* an error sits. We therefore resolve one Halo8 C–Br rearrangement frame by frame (Lee et al., 2025), fixed from reference structures alone before any prediction was inspected (Appendix C.6), against Linear, the reconstruction-only tokenizer (R-only), and Drift-React, the published generator with the best baseline TS-RMSD. Generation restores the given endpoints exactly, so all error is interior: in Figure 4b the outer columns coincide across rows while the boxed peak column does not, the bromine staying near its reactant carbon for Linear and R-only, partly transferred for Drift-React, and nearest the reference for REACFLOW. Both reference C–Br distances undergo 95% of their respective ~ 0.98 Å ranges between frames 1 and 3, whereas Linear spreads the change more gradually. At frame 4, breaking-bond extensions from frame 0 reach 30% of the reference extension for Linear, 21% for R-only, 53% for Drift-React and 69% for REACFLOW (Figure 4c). Per-frame aligned RMSD keeps that order at all eight internal frames, averaging 0.108 Å for REACFLOW against 0.244, 0.360 and 0.384 Å for Drift-React, Linear and R-only, respectively, with the largest absolute improvements near the reference peak (Figure 4d). Yet reference-path reconstruction remains comparable: 0.077 Å for R-only versus 0.082 Å for REACFLOW under the same metric (Appendix C). Better reconstruction thus need not yield better generation.

5 CONCLUSION

We present REACFLOW, a two-stage framework in which a frozen pathway tokenizer, pretrained with search-response supervision and endpoint-anchored interval completion, supplies the latent space for conditional flow matching. On Transition1x, clean reconstruction stays within 0.052–0.056 Å across pretraining objectives, yet REACFLOW improves endpoint-conditioned generation by 24.6–36.7% over reconstruction-only pretraining, with low mean geometry errors on both Transition1x and Halo8. Shaping the latent for transport, rather than reconstruction alone, is therefore a key lever; stochastic transport and coupling to electronic-structure refinement are future directions.

AI USE STATEMENT

In this work, we used generative AI tools only for language polishing, improving the grammar, wording, and clarity of text drafted by the authors. We have not used generative AI tools for research ideation, for the design of the proposed method or the experimental protocol, for implementation, for the generation or analysis of results, for the creation of figures and tables, or for literature discovery and citation; all of these are the original work of the authors. We have reviewed all AI-assisted work, checking every suggested revision against the authors' original text and the underlying results, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendix A document the datasets, realized partition sizes, and the reaction-level split rules, including reactant grouping for Halo8. The architecture and objectives are specified in Section 3 and Appendix B, with the complete schedules, loss weights, checkpoint rule, and hyperparameter table in Appendix D, and the matched settings shared by every controlled variant in Appendix C.3. Appendix C defines the metrics and evaluation conventions and Appendix C.1 the benchmark protocol; as noted in Appendix C.2, the controlled comparisons report one run per variant from a single seed. All baseline results in Table 1, including Drift-React, are taken from Schlama & Schwaller (2026); REACFLOW is evaluated using their released code under the same benchmark protocol. Full assumptions and proofs for the theoretical claims are given in Appendix E, and anonymized code, split manifests, and the scripts reproducing all tables and figures are provided in the supplementary materials.

REFERENCES

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 5817–5840. PMLR, 2024. URL https://proceedings.mlr.press/v235/castin24a.html.

Chenru Duan, Yuanqi Du, Haojun Jia, and Heather J. Kulik. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *arXiv preprint arXiv:2304.06174*, 2023. URL https://arxiv.org/abs/2304.06174.

Chenru Duan, Guan-Horng Liu, Yuanqi Du, Tianrong Chen, Qiyuan Zhao, Haojun Jia, Carla P. Gomes, Evangelos A. Theodorou, and Heather J. Kulik. React-OT: Optimal transport for generating transition states in chemical reactions. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-025-01010-0.

Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Physical Review B*, 66(5):052301, 2002. doi: 10.1103/PhysRevB.66.052301.

Kenichi Fukui. The path of chemical reactions—the IRC approach. *Accounts of Chemical Research*, 14(12):363–368, 1981. doi: 10.1021/ar00072a001.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://arxiv.org/abs/2111.06377.

Graeme Henkelman and Hannes Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of Chemical Physics*, 113(22):9978–9985, 2000. doi: 10.1063/1.1323224.

Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics*, 113(22):9901–9904, 2000. doi: 10.1063/1.1329672.

Aniketh Iyengar, Jiaqi Han, Pengwei Sun, Mingjian Jiang, Jianwen Xie, and Stefano Ermon. Align your structures: Generating trajectories with structure pretraining for molecular dynamics. In

International Conference on Learning Representations, 2026. URL https://arxiv.org/abs/2604.03911.

Hannes Jónsson, Greg Mills, and Karsten W. Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, pp. 385–404. World Scientific, 1998. doi: 10.1142/9789812839664_0016.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5562–5571. PMLR, 2021. URL https://proceedings.mlr.press/v139/kim21i.html.

Seonghwan Kim, Jeheon Woo, and Woo Youn Kim. Diffusion-based generative AI for exploring transition states from 2d molecular graphs. *Nature Communications*, 15:341, 2024. doi: 10.1038/s41467-023-44629-6.

Minhyeok Lee, Jinyoung Jeong, Islambek Ashyrmamatov, Umit V. Ucak, Sunwoo Kim, Juyong Lee, and Eunji Sim. A dataset of chemical reaction pathways incorporating halogen chemistry. *Scientific Data*, 12:1655, 2025. doi: 10.1038/s41597-025-05944-3.

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. URL https://arxiv.org/abs/2210.02747.

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://arxiv.org/abs/2209.03003.

Yufei Luo, Xiang Gu, and Jian Sun. Generating transition states of chemical reactions via distance-geometry-based flow matching. *arXiv preprint arXiv:2511.17229*, 2025. URL https://arxiv.org/abs/2511.17229.

Juno Nam, Miguel Steiner, Max Misterka, Soojung Yang, Avni Singhal, and Rafael Gómez-Bombarelli. Transferable learning of reaction pathways from geometric priors. *The Journal of Physical Chemistry Letters*, 16(45):11690–11699, 2025. doi: 10.1021/acs.jpclett.5c02620.

Zhanghan Ni, Yanjing Li, Zeju Qiu, Bernhard Schölkopf, Hongyu Guo, Weiyang Liu, and Shengchao Liu. Rigidity-aware geometric pretraining for protein design and conformational ensembles. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.02406.

Baron Peters, Andreas Heyden, Alexis T. Bell, and Arup Chakraborty. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *The Journal of Chemical Physics*, 120(17):7877–7886, 2004. doi: 10.1063/1.1691018.

Kalyan Ramakrishnan, Lars L. Schaaf, Chen Lin, Guangrun Wang, and Philip Torr. Implicit neural representations for chemical reaction paths. *The Journal of Chemical Physics*, 163(3):034109, 2025. doi: 10.1063/5.0267023. URL https://arxiv.org/abs/2502.15843.

Rémi Schlama and Philippe Schwaller. Drift-React: One-step generation of reaction pathways via SE(3) drifting fields. *arXiv preprint arXiv:2605.22990*, 2026. URL https://arxiv.org/abs/2605.22990.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x—a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9:779, 2022a. doi: 10.1038/s41597-022-01870-w.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Peter Bjørn Jørgensen, and Ole Winther. NeuralNEB—neural networks can find reaction paths fast. *Machine Learning: Science and Technology*, 3(4):045022, 2022b. doi: 10.1088/2632-2153/aca23e.

Ping Tuo, Jiale Chen, and Ju Li. Flow matching for reaction pathway generation. *arXiv preprint arXiv:2507.10530*, 2025. URL https://arxiv.org/abs/2507.10530.

Sifan Wang, Zhikai Wu, David van Dijk, and Lu Lu. GeoFunFlow: Geometric function flow matching for inverse operator learning over complex geometries. *arXiv preprint arXiv:2509.24117*, 2025. URL https://arxiv.org/abs/2509.24117.

Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good tokenizers. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1jBsi98fVe.

Yian Zhao, Feng Wang, Qiushan Guo, Chang Liu, Xiangyang Ji, Jian Zhang, and Jie Chen. Video generation with predictive latents. *arXiv preprint arXiv:2605.02134*, 2026. URL https://arxiv.org/abs/2605.02134.

A Dataset and Protocol Details

A.1 Dataset Construction and Partitions

Table 4 summarizes the reaction-level partitions. Both datasets use ten ordered frames with fixed atom correspondence: two endpoints and eight internal configurations. The tokenizer and Flow are trained from scratch separately for each dataset; Halo8 is an additional chemical-domain benchmark, not a zero-shot transfer experiment from Transition1x. All saved search states and augmentations inherit their parent reaction's partition, and normalization statistics are fitted on training reactions only.

Table 4: Dataset partitions used in this work. Counts refer to reactions, not frames or saved search states. Transition1x has no additional validation subset; its benchmark file named `valid_t1x` supplies the evaluation pool.

Dataset	Train	Validation	Evaluation	Elements
Transition1x	9,000	–	1,073	H, C, N, O
Halo8	15,348	1,934	1,891	H, C, N, O, F, S, Cl, Br

Transition1x. We recover native ten-frame reference pathways from the released Transition1x records and preserve the reaction-ID partition distributed with Drift-React: 9,000 training reactions and 1,073 evaluation reactions, with no reaction-ID overlap. We draw no additional validation subset or new random split. Molecules contain 4–23 atoms in training and 5–21 in evaluation, including hydrogens. Checkpoints are fixed by update count (Appendix D), not selected on this evaluation pool. Because the pool was inspected during development, we describe it as an evaluation set rather than a newly sealed test set.

Halo8. We construct the dataset from the official Halo8 ASE databases, following the split construction and evaluation protocol released with Drift-React. Of 19,176 reaction records, three have incomplete final pathways and are excluded, leaving 19,173 ten-frame references. Each final path combines the stored endpoints with the last complete eight-image internal band; no intermediate coordinates are resampled. We apply the released reactant-grouped, greedy 80/10/10 splitting rule with seed 42. Groups are keyed by canonical SMILES of the reactant connectivity inferred with RDKit; if connectivity perception fails, a key based on atomic numbers and a geometry hash is used. The resulting 1,927 groups (33 using the fallback) are kept intact, giving the counts in Table 4. Molecule sizes are 4–23, 6–18, and 6–19 atoms in training, validation, and evaluation, respectively. The Halo8 comparison in Table 1 therefore follows the same benchmark protocol as the reported baselines.

An earlier, family-stratified Halo8 test partition was inspected during development. It differs from the reactant-grouped benchmark split used for comparison with Drift-React (Schlama & Schwaller, 2026); no results or partition counts from that historical split are used in the current main tables. We disclose this prior exposure and do not treat any overlapping reaction as newly sealed evaluation data.

A.2 Recovering Labeled Search Pathways

For Transition1x, the original packing script stores ten images for the first retained pathway and only eight internal images for each later state (Schreiner et al., 2022a). For a reaction with n flattened configurations in this format, the number of retained states is $J = 1 + (n - 10)/8$. The first state is the first ten configurations. Each later state combines the original two endpoints with its corresponding eight-image internal block. We index positions, energies, and forces identically; reshaping the flattened array into ten-image blocks would misalign them.

This rule follows the packing format of the released Transition1x archive. The loader isolates records with incompatible lengths, identities, atom ordering, endpoints, or non-finite labels, and resolves remapped reaction identifiers through the release's index mapping. The clean target X is checked against the last saved state rather than replaced by a candidate. For Halo8, complete saved internal bands are joined with their reaction's endpoints using the database record order; incomplete bands are discarded. Positions, energies, and forces retain the same reaction, atom, and frame identities. Frame count is preserved instead of interpolating labeled geometries to fit the model.

Coordinates use the common translation defined by the reactant's geometric center, without frame-wise centering or rotation before label construction. Any common rotation or atom permutation is applied consistently to coordinates and physical forces. Labels are computed from the original paired structures in Å, eV, and eV/Å; perturbed or interpolated inputs do not inherit these labels. Each cache is keyed by dataset version, packing-code hash, unit conversion, and reaction mapping.

A.3 Label Eligibility and Information Boundaries

We fixed and recorded the tangent implementation, Cartesian metric, tie handling, peak-eligibility criteria, and numerical tolerances before running any evaluation. Tangent normalization and force projections exclude padded atoms and use a single $3N$-dimensional image vector. We used the physical force $-\nabla V$ recorded in the release, not a projected NEB or spring-augmented force.

The inherited candidate filter requires finite paired records, nondegenerate tangents and neighboring steps (norm above 10^{-10}), an internal peak at least 0.01 eV above both endpoints and 0.001 eV above its neighbors, minimum interatomic distance 0.55 Å, internal-response RMS in $[0.03, 5]$ eV/Å, and maximum atomic force norm at most 30 eV/Å. These are search-only training-data filters, not evaluation-reaction filters. At most eight eligible states per reaction are retained at approximately uniform positions in the saved sequence, followed by uniform within-reaction sampling. Some retained sets include a converged final state.

All tokenizer variants use the same retained candidate coordinates. The filter constants are inherited from the existing training protocol and fixed before comparisons. Climbing-stage metadata are not distributed with the release, so Equation 3 is a CI-inspired label rather than the recorded optimizer action. Response scales are fitted on training reactions only.

On the training partitions, Transition1x supplies 1,073,983 saved candidate states, of which 71,961 are retained across all 9,000 reactions after eligibility filtering and the eight-state cap. Halo8 supplies 2,007,307 saved-band candidates, including 40 incomplete bands that are discarded before eligibility filtering. In total, 122,710 bands are retained across 15,343 of the 15,348 Halo8 training reactions. The five reactions without eligible Y still contribute X/L reconstruction and completion but supply no candidate reconstruction or search loss. We substitute no interpolated or duplicated target for a labeled candidate. Each reconstruction component is averaged over the reactions for which it is defined, and the three component means then receive equal weight. This eligibility treatment is identical for all tokenizer controls, including R, so search-label availability does not change their coordinate-training sets.

Label validity is separate from structural validity. Structural masks describe real atoms and images and may be model inputs; energy-derived eligibility, peak indices, and response-validity masks are restricted to sampling and loss computation. Changing the stored labels with (Y, C) held fixed leaves the encoder, decoder, and auxiliary-head forward outputs unchanged.

B TOKENIZER ARCHITECTURE AND OBJECTIVES

Encoder and packing. For input W, let $Q_i = W_{:,i,:} \in \mathbb{R}^{T\times 3}$ denote atom i's trajectory, with $T = 10$ fixed for both datasets. The full-pathway encoder uses trajectory Gram features and invariant-coefficient vector synthesis, not a frame/temporal Transformer. Atom types use a 119-entry, eight-dimensional embedding followed by an $8 \to 128 \to 128 \to 32$ SiLU MLP. This type network is shared by the encoder, decoder, and search head. With $\Gamma(V) = \mathrm{vec}(VV^\top)/25$, the shared vector and scalar coefficient networks have widths $132 \to 128 \to 128 \to 40$ and $132 \to 128 \to 128 \to 4$. Their outputs yield $G = N^{-1}\sum_i A_i^g Q_i$ and $b = N^{-1}\sum_i b_i$, without softmax-normalized pooling weights. The local coefficient network has widths $232 \to 128 \to 128 \to 56$ and a zero-initialized final layer. The fixed Φ is obtained by QR orthogonalization of constant through cubic modes on the ten-point grid, with a fixed sign convention.

The local network reads $[h_i, \Gamma([Q_i; G]), b]$ and outputs invariant coefficients $A_i \in \mathbb{R}^{K\times(T+K)}$. The complete atom-wise latent is

$$Z_i = T^{-1}\Phi^\top Q_i + \tfrac{1}{4}A_i[Q_i; G], \qquad \Phi^\top\Phi = TI_K. \tag{10}$$

The first term is a fixed temporal projection and the second a learned geometric correction, both stored in Z_i as compressed coefficients rather than raw coordinates. Zero initialization of the correction leaves the four-mode projection as the initial local encoder.

For a padded batch, the latent has shape $B \times (N_{\max}+1) \times 4 \times 4$. Token zero contains (G, b); atom tokens contain $(Z_i, 0)$. Each last-axis tuple represents three vector components and one scalar, not a four-dimensional geometric vector. There are $12N + 16$ active scalar entries. Padding atoms and the atom tokens' scalar slots are structural zeros, excluded from latent statistics and velocity losses.

Decoder and task mode. The decoder synthesizes full coordinates from $\overline{V}_i = [Z_i; G]$:

$$D^\rho_\psi(z)_{k,i} = \sum_{\ell=1}^{K}\Phi_{k\ell}Z_{i,\ell} + \tfrac{1}{4}\sum_{\ell=1}^{2K}\alpha_{k,i,\ell}(z, a, m, s_k, \rho)\overline{V}_{i,\ell}. \tag{11}$$

Here $\Phi \in \mathbb{R}^{T\times K}$ is the fixed temporal basis with $\Phi^\top\Phi = TI_K$. Frames share weights and are decoded in parallel. Common translation is handled by the reactant-centered coordinate frame.

The decoder forms invariant features $d_{k,i} = [h_i, \Gamma([Z_i; G]), b, \eta(s_k)] \in \mathbb{R}^{117}$, where η contains s_k and eight sine/cosine pairs. The task flag enters through

$$h^{(0)}_{k,i} = W_d d_{k,i} + b_d + \rho w_\rho, \qquad w_\rho \in \mathbb{R}^{256}. \tag{12}$$

The vector w_ρ is zero-initialized. This is equivalent to a 118-input projection but preserves the original 117-input computation and initialization, adding only 256 parameters. Four pre-norm Transformer layers use width 256, eight heads, FFN width 1024, GELU, and zero dropout. The four layers have independent parameters; the same decoder stack is applied to every frame. Frames are folded into the batch dimension, and attention exchanges information across atoms within each frame. A zero-initialized $256 \to 8$ readout supplies the coefficients in Equation 11. Raw decoding reconstructs all ten frames; it receives neither R/P coordinates nor L. The task flag changes the shared decoder's behavior, not the latent layout.

Search readout. For atom i, the head uses $V_i^H = [Z_i; G; R_i; P_i] \in \mathbb{R}^{10\times 3}$ and features $[h_i, b, \mathrm{vec}(V_i^H(V_i^H)^\top)] \in \mathbb{R}^{136}$. Unlike the tokenizer's Gram features, the head Gram matrix is not divided by 25. A $136 \to 64 \to 80$ network with SiLU predicts 8×10 coefficients for synthesizing the eight internal response vectors. Endpoint outputs are zero. Uncompressed internal coordinates, tangent estimates, and peak indices do not enter the head.

Loss boundaries. Equation 5 gives equal weight to $X/L/Y$ reconstruction, normalized by $3TN$ within each reaction. Reconstruction-only latent noise has standard deviation 0.01; ignored structural slots remain irrelevant to decoding. Search and completion use clean latents. The response RMS is estimated by averaging within candidates and then equally across reactions, with a floor of 10^{-6}. All internal images receive equal search-loss weight. The ordinary-force control fits its own training RMS; the common response probes retain one shared $\mathcal{U}$ scale.

Interval completion samples the length uniformly from one to eight and the valid start uniformly, retaining both endpoints. Every input-dependent feature is recomputed after interpolation replace-

ment. The masked and retained terms are normalized separately and weighted equally, including the retained endpoints. Completion supplies no energy or force labels.

Flow endpoint conditioning. The Flow has its own trainable endpoint conditioner, independent of the frozen tokenizer, receiving only gauge-fixed reactant and product coordinates, atomic numbers, and the atom mask. Its separate 119-entry element embedding has width eight. For each atom pair (i, j), the two element embeddings are summed and concatenated with three invariant geometric features, $[\|R_i - R_j\|^2, \|P_i - P_j\|^2, (R_i - R_j)^\top (P_i - P_j)]/25$. An $11 \to 128 \to 128 \to 32$ SiLU MLP embeds these pair features, followed by mean pooling over valid neighbors. A second $44 \to 128 \to 128 \to 32$ SiLU MLP combines the pooled 32-dimensional representation, the eight-dimensional element embedding, and the four entries of the 2×2 endpoint-vector Gram matrix. This yields a 32-dimensional context per atom; masked mean pooling gives the additional shared token's context. The conditioner receives no internal reference frames, energies, forces, or reference peak index. It belongs only to the Flow: the AE decoder does not directly read endpoint coordinates or conditioner features.

Velocity parameterization. At transport time τ, the evolving code z_τ and fixed source $z_s = E_\phi(L)$ are normalized by training-set latent scales S. Each token's scalar input concatenates its 32-dimensional endpoint context, τ and τ^2, sine/cosine features at six dyadic frequencies, a two-dimensional shared/atom indicator, and the four scalar channels of each of z_τ/S and z_s/S. The time features are $[\tau, \tau^2, \{\sin(2\pi 2^j \tau), \cos(2\pi 2^j \tau)\}_{j=0}^{5}]$, giving $32+14+2+4+4 = 56$ scalar inputs. The vector input contains four channels from each normalized code and four endpoint channels, giving 12 three-dimensional vectors. The endpoint channels are $[R_i, P_i, R_i, P_i]$ for atom tokens and the corresponding masked means for the shared token. A scalar–vector embedding maps these inputs to 256 invariant scalar and 48 equivariant vector channels. Six directional pair-attention blocks with four heads operate over the $N + 1$ valid tokens. Reactant and product pair distances each use 16 radial basis functions over 0–8 Å (32 features in total); endpoint-relative directions enter through equivariant vector operations and invariant contractions.

The velocity readout uses a bias-free $48 \to 4$ channel projection for the vector output, modulated by a sigmoid gate predicted from the 256 scalar channels. Each gate is shared across the three spatial components of its vector channel. A separate $256 \to 4$ linear projection produces scalar velocities. The outputs are packed into the original $(N + 1) \times 4 \times 4$ layout, masked to remove padded atoms and atom-token scalar slots, and rescaled to the original latent units. Inference starts from the actual $E_\phi(L)$, not Gaussian noise or a zero code, and uses two Heun steps (NFE $= 4$). The terminal code is decoded in mode $\rho = 0$; reactant-centering is then undone and the known endpoints restored outside the decoder.

Latent normalization.

Scales are fitted from training clean-target latents, separately for local vectors, shared vectors, and shared scalars. Vector xyz components share one RMS scale per channel; no mean subtraction or per-axis whitening is applied. Local statistics average atoms within each reaction before averaging reactions, and every active scale has a floor of 10^{-5}. Internal Flow features use z/S, and predicted velocities are multiplied back by S, leaving the integration state in the original latent units. For scaled velocity error e, the branch norm is

$$\|e\|_{\mathrm{br}}^2 = \frac{\|e_0\|_F^2}{2(4K)} + \frac{\sum_{i=1}^{N} \|e_i^{\mathrm{vec}}\|_F^2}{2(3NK)}. \tag{13}$$

Every tokenizer variant therefore received fresh caches, fresh scales, and a freshly trained Flow, even when its packed latent had the same width.

C Representation Diagnostics and Metrics

C.1 Benchmark Sources and Metric Conventions

Source and scope. Table 1 compares our evaluation of ReacFlow with the baseline results reported in Table 1 of Schlama & Schwaller (2026) (arXiv:2605.22990v1). All baseline means and standard deviations—Linear, IDPP, Geodesic, NeuralNEB, and Drift-React—are taken directly from that work; no baselines are retrained or re-evaluated for this table. On both Transition1x and Halo8,

we evaluate REACFLOW using the released Drift-React code under the same geometry and energy evaluation protocol. The NeuralNEB results correspond to the default PaiNN/CI-NEB retraining reported in that work. The reported spreads are standard deviations, not uncertainty across our pipeline seeds. TS-only results for ReactOT, OA-ReactDiff, and TSDiff are excluded from the full-pathway table because they do not supply the same multi-image output. MEPIN requires a separate matched-level, matched-split comparison.

Reporting convention. Generation geometry evaluated in this work, including the ablations in Tables 2 and 3, follows the released Drift-React evaluation protocol and uses the N-normalized scale, with point distance $d_N = \sqrt{\sum_{i,c} \Delta_{ic}^2 / N}$. The generated IRC-RMSD in Figure 3a,b follows the same protocol. The reconstruction values in Table 2 and Figure 3a use the same N normalization but the separate raw, all-frame reconstruction definition below. The relation $d_N = \sqrt{3}\, d_{3N}$ converts per-coordinate RMSD to this reporting scale; it changes only the denominator, not peak selection, alignment, or frame aggregation. Training losses and theoretical risks retain their per-coordinate conventions, with their relation to reported RMSDs specified below.

Geometry and energy metrics. For REACFLOW in Table 1 and the controlled variants in Tables 2 and 3, we adopt the TS-frame selection implemented in the released Drift-React evaluator: the reference-energy peak $k_{\mathrm{ref}} = \arg\max_k E_{\mathrm{ref}}(k)$ over all ten frames, choosing the earliest frame in a tie. TS-RMSD compares the generated and reference structures at this shared index; predicted energies are not used to select the evaluated geometry. All three geometry metrics retain the released evaluator's metric-specific alignment and valid-atom handling, with fixed atom correspondence and atom-count (N) normalization. IRC-RMSD averages corresponding-frame errors over all ten frames, including endpoints, then equally over reactions. Discrete Fréchet permits order-preserving correspondence over the complete paths and measures the worst matched distance. These benchmark definitions are separate from the proper-rotation ($\mathrm{SO}(3)$), eight-internal-frame Path diagnostic below. The two energy columns use the reference-peak barrier and corresponding-frame errors defined below.

We retain both energy columns because a correct barrier does not imply a correct energy profile. Conversely, geometric similarity and even low single-point energy error do not certify stationarity, a first-order saddle, or endpoint connectivity. IRC-max and raw-coordinate error remain useful internal diagnostics, but adding them to the external table would require new baseline evaluations. Success rates and refinement cost likewise require a common convergence and connectivity protocol, not inferred numbers.

Protocol conventions. For both Transition1x and Halo8, our evaluation follows the benchmark protocol implemented in the released Drift-React code. Both datasets use ten images: two endpoints and eight internal frames. Halo8 uses the released reactant-grouped split procedure. Dataset versions, partitions, and prior data exposure are described in Appendix A.

Single-point energy evaluation. The generated REACFLOW pathways are scored using the released Drift-React energy evaluator under the same benchmark protocol. The dataset-specific electronic-structure settings used for our evaluation are given below. Generated geometries are scored directly, with identical atom ordering and frame correspondence to the references and padded atoms removed. No geometry optimization, transition-state refinement, force evaluation, or Hessian calculation is performed. Transition1x uses PySCF 2.14.0 RKS at ωB97x/6-31G(d) (6-31G*), with SCF tolerance 10^{-6}, at most 200 iterations, the default numerical integration grid, and no density fitting. Halo8 single-point calculations use ORCA 6.0.1 at ωB97X-3c with total charge 0 and spin multiplicity 1. Predicted energies are compared against the frame-matched reference energies retained from the dataset records; those reference energies are not recomputed. Hartree energies are converted using $1\,\mathrm{Ha} = 27.2114$ eV.

For each reaction, the reference-peak barrier error is

$$\left|\Delta E^{\ddagger}\right| = \left|[E_{\mathrm{pred}}(k_{\mathrm{ref}}) - E_{\mathrm{pred}}(0)] - [E_{\mathrm{ref}}(k_{\mathrm{ref}}) - E_{\mathrm{ref}}(0)]\right|. \tag{14}$$

Image-energy MAE first averages $|E_{\mathrm{pred}}(k) - E_{\mathrm{ref}}(k)|$ over valid paired frames within each reaction, then equally over reactions with at least one valid pair. Barrier MAE instead averages over

reactions with valid pairs at both frame 0 and k_{ref}; these two metrics can therefore have different valid reaction sets. A predicted energy is valid only if its calculation converges successfully and returns a finite value, and its paired reference energy must also be finite. Failed or missing frames are never assigned zero or interpolated, and a failed frame does not discard the reaction's remaining valid image-energy pairs. Counts of attempted, successful, failed, and missing calculations and the valid denominators are retained with the aggregates. Energy postprocessing follows the released Drift-React benchmark protocol, including MAD filtering ($k = 3$) for Transition1x image-energy summaries.

C.2 Reconstruction and Auxiliary Representation Diagnostics

The generation columns of Tables 1–3 and the generated IRC-RMSD in Figure 3a,b follow the benchmark protocol in Appendix C.1. The proper-rotation distances below instead define auxiliary and case-level diagnostics; their eight-internal-frame Path average is not the benchmark IRC-RMSD.

Configuration and path errors. For images $A, B \in \mathbb{R}^{N\times 3}$ with fixed atom correspondence and N valid atoms, define the N-normalized reporting distances

$$d_{\mathrm{raw}}(A,B) = \frac{\|A-B\|_F}{\sqrt{N}}, \qquad d_{\mathrm{align}}(A,B) = \min_{\substack{O\in \mathrm{SO}(3)\\ c\in\mathbb{R}^3}} \frac{\|AO + \mathbf{1}c^\top - B\|_F}{\sqrt{N}}. \tag{15}$$

Path and Raw Path average their respective distances over the eight internal frames, then over reactions. Raw Path compares structures in the common endpoint frame, not independently recentered images. $\mathrm{Peak}_{\mathrm{ref}}$ uses d_{align} at the reference maximum-energy internal index. An auxiliary proper-rotation Fréchet diagnostic applies the standard ordered discrete-path recurrence to all ten frames using d_{align}; it does not redefine the benchmark Fréchet columns. Independent alignment measures configuration agreement modulo rigid motion; Raw Path also detects inconsistent frame-to-frame rigid displacements.

Clean reconstruction averages $d_{\mathrm{raw}}(D^0_\psi(E_\phi(X))_k, X_k)$ over all ten frames before endpoint replacement. Table 2 and Figure 3a report this N-normalized RMSD; per-coordinate ($3N$-normalized) reconstruction values are multiplied by $\sqrt{3}$ before rounding. Clean peak reconstruction instead evaluates d_{align} at the same reference energy-maximum internal index used for generated $\mathrm{Peak}_{\mathrm{ref}}$, with identical atom masks and reaction averaging. It checks clean-target peak reconstructability, not decoder sensitivity to off-target predicted latents. Source and held-out-candidate reconstruction are separate diagnostics. Neither clean reconstruction nor a difference of reported means is treated as a pure latent-transport error.

For the single-reaction comparison in Section 4.4, each variant reconstructs the same clean reference path using its own frozen encoder–decoder. Reconstruction and generation RMSDs both average d_{align} over the eight internal frames with the same valid-atom mask. These case-level reconstruction values therefore differ in alignment and frame aggregation from the raw all-frame reconstruction metric above.

Decoder-free terminal-latent diagnostics. For each frozen tokenizer, let $\widehat{z}_q(C)$ be the actual two-Heun-step output of the Flow checkpoint after q updates. Here *terminal* refers to the end of Flow integration; the latent represents the complete pathway, not its final molecular frame. Starting from $z_s = E_\phi(L)$, the rollout uses only legal source and endpoint inputs; $z_t = E_\phi(X)$ is never used to construct an intermediate rollout state. With the same train-only S and branch norm as in Equation 8, we compute

$$A_q = \mathbb{E}\,\|(\widehat{z}_q - z_t)/S\|^2_{\mathrm{br}}, \qquad B = \mathbb{E}\,\|(z_s - z_t)/S\|^2_{\mathrm{br}}, \qquad r_q = A_q/B. \tag{16}$$

The scaled error also equals the error of the predicted displacement $\widehat{z}_q - z_s$ against $z_t - z_s$. Each expectation averages the common evaluation reactions for the fixed run; r_q is the ratio of these aggregate squared errors, not an average of per-reaction ratios. Figure 3c reports r_q at the final checkpoint. If $B \le 10^{-12}$, the ratio is undefined and A_q, B are reported separately. A ratio below one indicates improvement over leaving the source latent unchanged; it is not an explained-variance score or a common absolute metric across tokenizers.

The FM-only checkpoint is fixed immediately after 4,096 updates and before geometry-supervised training; the final checkpoint is the same 20,480-update checkpoint used for generated-path scoring.

No latent-error-based checkpoint selection is performed, and the diagnostic reuses the saved checkpoints and latent caches. Because the scores use a train-scaled branch norm, they remain tied to the tokenizer's parameterization: they measure transport progress within a fixed latent space rather than intrinsic latent complexity, and they are not the unscaled latent norm that appears in the theorems.

Per-case re-encoded path discrepancy. For Figure 5, each complete Reference, Linear, Drift-React, R-only, and Ours (REACFLOW) coordinate path is independently passed through the same frozen Full (R+S+C) encoder for its dataset. For case c, method m, dataset d, and training-only channel scales S_d, we define

$$D_z(c,m) = \left\| \frac{E^{\mathrm{Full}}_{\phi,d}(X_{c,m}) - E^{\mathrm{Full}}_{\phi,d}(X_{c,\mathrm{ref}})}{S_d} \right\|^2_{\mathrm{br}}. \tag{17}$$

The same atom mask and active shared/local scalar/vector channels are used within a case, with structural padding excluded as in Equation 8. Thus $D_z(c,\mathrm{ref}) = 0$; the quantity is an absolute latent distance, so the Linear path is not normalized to one. This re-encoded full-path diagnostic is distinct from Figure 3c, which scores a Flow terminal latent in each tokenizer's own latent space.

Reported metrics and theoretical risks. Clean reconstruction and the separate Path/Raw Path diagnostics average N-normalized per-image RMSD, whereas Theorem 3 retains per-coordinate MSE, normalized by $3NT$ over a complete path. The training losses and theoretical norm conventions are unchanged. Accounting for this factor of three, Jensen's inequality gives, for the same reactions and predictions,

$$\mathrm{Recon}_{\mathrm{mean}} \le \sqrt{3\mathcal{R}_{\mathrm{rec,clean}}}, \qquad \mathrm{Path}_{\mathrm{aligned}} \le \mathrm{RawPath} \le \sqrt{\frac{3T}{T-2}}\sqrt{\mathcal{R}_{\mathrm{gen}}}. \tag{18}$$

Here $\mathrm{Path}_{\mathrm{aligned}}$ denotes the proper-rotation, eight-internal-frame diagnostic, not benchmark IRC-RMSD. The latter inequalities also use optimal alignment and the fixed generated endpoints. Mean reconstruction RMSD therefore cannot replace the reconstruction-risk term on the right of the theorem; we do not evaluate the bound numerically, which would require the second-moment risks.

Tables 2 and 3 report one complete tokenizer-to-Flow run per controlled variant, evaluated on the same reactions, with all frames and candidates from a reaction kept together when computing its score. Their entries are reaction-level point estimates from a single training seed, so we report no between-run variance, confidence interval, or significance claim, and the differences between means are not evidence of robustness to initialization.

C.3 MECHANISM CONTROLS

Force-supervision control. The physical-force target is the recorded F_k at each internal image, with zero endpoint targets. This control uses the same eligible candidates, head architecture, search-weight schedule, and completion objective as R+S+C. The target is normalized by its training RMS so that its magnitude alone does not change the effective loss scale.

Completion gradients into the encoder. The stop-E control uses the same architecture, initialization seed, $X/L/Y$ sampling, masks, weights, and update schedule as R+S+C. For the completion branch alone, the latent $E_\phi(\widetilde{X})$ is detached, as are the shared atom-type features read by the decoder. More generally, completion contributes no gradient to the encoder or to parameters shared with it, updating only decoder-exclusive parameters; the encoder is still trained, since reconstruction and search retain their normal gradients.

A completion-only backward pass checks that encoder and shared-type gradients are zero while decoder-exclusive gradients remain intact. Since decoder updates can still affect subsequent reconstruction gradients, this control tests the direct completion-gradient route, not complete causal isolation. Each resulting tokenizer is frozen, recached, and followed by fresh Flow training.

Masking structure. The random-frame control first draws the same hidden-frame count as interval completion, then selects that many internal positions uniformly without replacement. Each missing run is interpolated between its nearest retained boundaries. Endpoints, full-mask probability, decoder mode, loss weights, and budget are unchanged. The control is matched on hidden-frame count by design; scattering those frames still changes how much context surrounds each one.

C.4 Latent Transport Improves upon Direct Completion

Question. Full internal masking makes the completion input exactly L, so the frozen completion mode already defines an endpoint-only pathway generator. We compare this existing predictor with the full ReacFlow generation procedure on Transition1x to assess the additional benefit of learned latent transport.

Matched tokenizer and inference routes. We use the same frozen Transition1x R+S+C tokenizer as in Table 1 and its associated final Flow checkpoint, without additional training or parameter updates. Starting from the same $z_s = E_\phi(L(C))$, the two routes produce

$$\begin{aligned} \widehat{X}_{\mathrm{comp}} &= P_C D^1_\psi(z_s) = Q(C), \\ \widehat{X}_{\mathrm{flow}} &= P_C D^0_\psi\big(\widehat{z}(C)\big), \end{aligned} \tag{19}$$

where $\widehat{z}(C)$ is the Flow output obtained from z_s by two Heun steps with four velocity evaluations. Direct completion is precisely the coordinate predictor used in Theorem 1, not target reconstruction $D^0_\psi(E_\phi(X))$. These are two independent generation routes: the Flow starts from z_s, not from the completion output.

Information and preprocessing controls. Both learned routes use the same 1,073 evaluation reactions, ten frame positions, atom ordering, valid-atom masks, and reactant-centering convention. We construct L from the endpoints and recompute all encoder features from L; reference internal coordinates, energies, and peak indices enter only scoring. All modules remain frozen and run in evaluation mode without latent noise. The completion flag enters only the decoder, whose latent inputs retain their original units rather than being divided by the Flow scale S. Both routes restore the common translation and known endpoints identically, without the search head, physical refinement, or electronic-structure evaluations during generation.

Two latent predictors in one frozen space. Let $z_t = E_\phi(X)$ and $g(C) = E_\phi(Q(C))$. Using the same training-set channel scales S, active channels, and branch norm as in Equation 16, define

$$\begin{aligned} A(h) &= \mathbb{E}\,\|(h(C) - z_t)/S\|^2_{\mathrm{br}}\,, \qquad B = A(z_s), \\ r_{\mathrm{comp}} &= A(g)/B, \qquad r_{\mathrm{flow}} = A(\widehat{z})/B, \qquad r_{\mathrm{source}} = 1. \end{aligned} \tag{20}$$

Expectations average the same evaluation reactions; these are ratios of aggregate squared errors, not averages of per-reaction ratios. If $B \le 10^{-12}$, report the numerator and denominator instead. The completion predictor is decoded, endpoint-restored, and re-encoded, whereas the Flow predictor is its terminal latent before final decoding. Thus these scores compare two latent predictors in a shared representation, not identically re-encoded output pathways or successive stages of one inference chain.

Metrics and reporting. We evaluate both learned routes with the same released Drift-React geometry evaluator as Table 1 (Appendix C.1), including its reference-energy peak selection, atom-count normalization, all-ten-frame IRC averaging, and metric-specific alignment. Table 5 reports mean geometry errors across evaluation reactions and the aggregate latent ratios. These are descriptive comparisons of fixed checkpoints, not estimates of variability across training runs. For context, the Linear geometry means are quoted from Table 1, rather than recomputed for this control. Linear's latent ratio is one by definition, not a literature measurement. The comparison of interest is between the two learned routes on identical evaluation reactions.

Results and interpretation. Direct completion produces a useful endpoint-conditioned prediction: its re-encoded latent has a relative discrepancy of 0.472, compared with the source value of 1. The full transport route achieves lower mean errors on all three geometry metrics. Relative to direct completion, ReacFlow reduces TS-RMSD from 0.249 to 0.215 Å (13.7%), IRC-RMSD from 0.201 to 0.155 Å (22.9%), and Fréchet distance from 0.333 to 0.260 Å (21.9%). Its terminal latent also has a lower relative discrepancy, 0.314, than the re-encoded completion prediction. This is a comparison between two latent predictors, not a sequential reduction of a completion residual by the Flow.

Table 5: Direct completion versus latent transport on Transition1x. Both learned routes use the same frozen R+S+C tokenizer. Geometry columns report means in Å; lower is better. Linear geometry (†) is quoted from Table 1 (Schlama & Schwaller, 2026). REACFLOW uses the geometry means of Table 1 and the final latent ratio of Figure 3c.

Endpoint-conditioned route	Latent ratio	TS-RMSD	IRC-RMSD	Fréchet
Linear interpolation	1.000	$0.432^{\dagger}$	$0.364^{\dagger}$	$0.748^{\dagger}$
Direct completion	0.472	0.249	0.201	0.333
REACFLOW	0.314	0.215	0.155	0.260

Latent ratios are dimensionless. Completion uses a re-encoded pathway latent, while REACFLOW uses its terminal Flow latent, as defined in Equation 20.

These results support the added value of the trained latent-transport route beyond the existing completion branch. They complement Theorem 1, which bounds the error of a completion-induced latent predictor but does not rank it against the learned Flow. The comparison fixes the tokenizer while changing the decoder mode and use of the additionally trained Flow; it does not establish equal-total-budget superiority or the necessity of flow matching over other endpoint-conditioned predictors.

C.5 QUALITATIVE PATHWAY COMPARISONS

Figure 5 compares nine reference-selected reactions grouped by H transfer, heavy-atom rearrangement, and halogen-containing rearrangement. All methods receive identical endpoints; structures are shown at the reference-defined peak frame and the latent discrepancy is computed over the complete path. The case IDs were fixed from reference structures alone, before any prediction was inspected (Appendix C.6). These nine cases illustrate sample-level variation; the aggregate comparison remains Table 1.

C.6 REFERENCE-ONLY CASE SELECTION

Eligibility and chemical motifs. We selected the nine reactions in Figure 5 using only held-out reference structures and energies, before loading model checkpoints, predictions, or errors. Endpoint connectivity is screened using distances below 1.25 times the sum of ASE covalent radii. H transfer requires a hydrogen to lose one heavy-atom neighbor and gain another without changing heavy–heavy connectivity. Heavy-atom rearrangement requires at least one broken and one formed heavy–heavy connection, excluding changes to C–halogen connections. Halogen rearrangement requires a changed C–F, C–Cl, C–Br, or C–I connection. These are geometric selection motifs, not verified bond orders or reaction mechanisms. Eligible references contain 8–20 atoms, finite coordinates and energies, an internal reference-energy peak, and no heavy-atom separation below 0.6 Å anywhere along the path.

Reference-only complexity score. For each eligible reaction, the selection heuristic is

$$s_{\text{case}} = 3n_{\Delta\text{HH}} + \min(n_{\text{move}}, 5) + 2\frac{\delta_{\text{NL}}}{1\,\text{Å}}, \qquad \delta_{\text{NL}} = \max_{k,\, i\in\mathcal{H}} \|X_{k,i} - L_{k,i}\|_2, \tag{21}$$

where $\mathcal{H}$ is the set of non-hydrogen atoms and $n_{\Delta\text{HH}}$ counts broken plus formed heavy–heavy endpoint connections. To compute n_{move}, we align the product to the reactant once by a fixed-correspondence, proper-rotation Kabsch fit on heavy atoms and count displacements above 0.5 Å. In contrast, δ_{NL} uses the reference path and its endpoint-linear interpolation in their common coordinate frame, without independent framewise alignment. It is a maximum atomic displacement, not a path RMSD. The score is used only to select examples, not to evaluate a model.

Deterministic choice of three cases per motif. We retain one previously fixed reference-only representative per motif: Transition1x `rxn0170`, Transition1x `rxn9094`, and Halo8 `Halogen_BrC5H6NS_rxn11512`. For the two additional examples, we first restrict each motif's eligible pool to its 50th–85th percentile score band. Within that band, we select unused candidates nearest its own 50th and 70th percentile scores, breaking ties by score, dataset, and reaction ID.

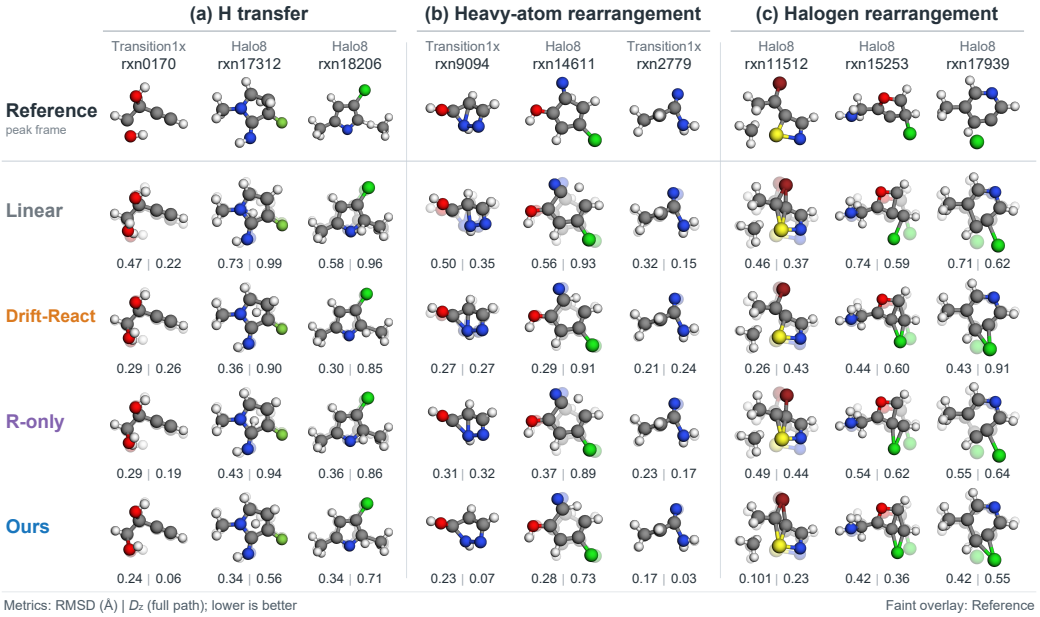

Figure 5: Across H transfer, heavy-atom rearrangement, and halogen-containing rearrangement, Ours (REACFLOW) tracks the reference peak-frame geometry more closely than Linear or R-only in these nine examples. Rows compare Reference, Linear, Drift-React, R-only, and Ours under shared endpoint conditioning at the reference-defined peak frame. R-only uses the reconstruction-only objective of Table 2, trained separately on each dataset. Faint overlays show the reference. Beneath each prediction, the left value is RMSD (Å) and the right is the full-path re-encoded latent discrepancy D_z (Appendix C.2). The displayed configurations are not certified transition states.

Thus the percentile targets are computed within the restricted band; the previously fixed case need not lie in that band. The eligible pools contain 616 H-transfer, 907 heavy-atom, and 159 halogen-rearrangement reactions. Table 6 lists the resulting IDs. All methods are evaluated on these same reactions, with no subsequent selection based on prediction quality.

Table 6: Reference-selected cases, in the left-to-right order used within each motif in Figure 5. Halo8 reaction IDs omit the common `Halogen_` prefix. The first case in each motif is the previously fixed representative.

Motif	Dataset	Reaction ID
H transfer	Transition1x	`rxn0170`
	Halo8	`C5FH7N2_rxn17312`
	Halo8	`C6ClH8N_rxn18206`
Heavy-atom rearrangement	Transition1x	`rxn9094`
	Halo8	`C5ClH4NO_rxn14611`
	Transition1x	`rxn2779`
Halogen rearrangement	Halo8	`BrC5H6NS_rxn11512`
	Halo8	`C5ClH6NO_rxn15253`
	Halo8	`C6ClH6N_rxn17939`

D OPTIMIZATION AND CHECKPOINT PROTOCOL

Runtime and random seeds. Training uses a single NVIDIA GeForce RTX 5090 GPU in one process, with PyTorch 2.11.0 and CUDA 12.8. Model computation uses float32 without automatic mixed precision; float32 matrix-multiplication precision is set to `highest`. The fixed random seeds are 20261201 for tokenizer initialization, 20261202 for batch sampling, 20261203 for search-

candidate sampling, 20261204 for completion intervals, 20261110 for Flow initialization, 20261111 for Flow-time sampling, and 20260911 for the reversal-augmented phase. These identify stochastic components of one run, not independent repeated trials. Transition1x and Halo8 use separate model weights, training caches, and normalization statistics.

Pretraining and Flow schedules. After the FM-only warm-up, the Flow objective combines aligned and raw geometry with flow matching:

$$\mathcal{L}_{\mathrm{Flow}} = 0.25\,\mathcal{L}_{\mathrm{SO(3)}} + 0.10\,\mathcal{L}_{\mathrm{raw}} + 0.01\,\mathcal{L}_{\mathrm{FM}}. \tag{22}$$

The geometry terms average framewise $3N$-normalized RMSD with and without proper-rotation alignment. Differentiable Heun rollouts start from the actual source latent and decode through frozen D_ψ^0; endpoints are restored outside the decoder.

Table 7 records the documented full-pathway schedule, including interval completion, and the settings inherited by the controlled comparisons. With zero-based tokenizer update q, search supervision uses

$$\lambda_s(q) = 10^{-3}\,\mathrm{clip}\big((q-1024)/1024, 0, 1\big). \tag{23}$$

Only the search objective is delayed: the documented completion run uses $\lambda_c = 1$ from the first tokenizer update. This setting is fixed rather than selected using the final evaluation pool, and is retained for the matched masking and gradient controls. The search ramp is distinct from the shorter learning-rate warm-up. The no-completion control sets $\lambda_c = 0$, leaving the search schedule and the $X/L/Y$ self-reconstruction losses unchanged.

Table 7: Training schedule for the full-pathway model with interval completion. These settings produce the REACFLOW (R+S+C) rows of Table 1. Table 2 varies the pretraining-objective combination; Table 3 replaces one component of R+S+C at a time. Other settings remain fixed.

Setting	Value
Path frames (both datasets)	10 (2 endpoints + 8 internal)
Tokenizer updates	8,192
Flow updates (total / final reversal-augmented)	20,480 / 4,096
FM-only updates	4,096
Batch size	32
Optimizer / weight decay	AdamW / 10^{-4}
Tokenizer and main Flow peak LR	3×10^{-4}
Reversal-augmented phase peak LR	3×10^{-5}
LR warm-up / final-to-peak ratio	256 updates / 0.1
Gradient-norm clip	1
Reconstruction latent-noise standard deviation	0.01
Completion weight / activation	1 / first tokenizer update
Geometry-training Heun steps	Alternating 2 and 4
Inference Heun steps / NFE	2 / 4

Each optimization stage starts a new optimizer; the learning rate decays by cosine after warm-up, and incomplete final batches are retained. Geometry training uses Equation 22, averaging all-frame errors after endpoint replacement. Proper-rotation alignment transforms are held constant during differentiation; gradients through the aligned coordinate error and frozen decoder still reach the Flow. A two-step training rollout costs four velocity evaluations and a four-step rollout costs eight, each in addition to the FM call. Inference NFE is listed separately in Table 7. Within the geometry-supervised phase, the final 4,096 updates reverse reactions with probability 0.5, recenter and re-encode both pathways, and keep the original training-latent scales fixed. These updates are included in the total of 20,480 Flow updates. The inference rule remains two Heun steps without source noise or the search head.

Checkpoint and comparison scope. The controlled tokenizer variants use fixed 8,192-update checkpoints and freshly initialized Flows evaluated after 20,480 updates; the 4,096-update checkpoint is used only for the FM-only diagnostic. Each controlled ablation in Tables 2 and 3 reports one complete run. Table 1 compares REACFLOW, evaluated using the released Drift-React code under the same benchmark protocol, with the reported results for Linear, IDPP, Geodesic, NeuralNEB, and Drift-React from Schlama & Schwaller (2026). This benchmark comparison does not match training

budgets or establish relative compute efficiency. Training manifests, candidate caches, normalization statistics, run and checkpoint identifiers, and hyperparameters specify our models' reproducible evaluation.

E Analysis of Search-Aware and Predictive Tokenization

These results characterize supervision, endpoint-conditioned latent prediction, and error propagation for fixed parameters and normalization statistics. Search expectations use reaction-balanced eligible candidates; completion and generation use a common target-reaction distribution. The statements apply to either empirical or population risks under their stated conditions. Applying them to an empirical distribution does not establish a test-distribution guarantee.

Norms and conventions. Padding is excluded. For $\mathcal{A} \subseteq \{0, \dots, T-1\}$ with $|\mathcal{A}| > 0$, define

$$\|W\|_{\mathcal{A}}^2 = \frac{1}{3N|\mathcal{A}|} \sum_{k \in \mathcal{A}} \sum_{i=1}^{N} \|W_{k,i}\|_2^2, \qquad \ell_{\mathcal{A}}(\widehat{W}, W) = \|\widehat{W} - W\|_{\mathcal{A}}^2. \tag{24}$$

The subscripts int and all select internal and all frames, respectively. For the code in Equation 2,

$$\|z\|_z^2 = \|G\|_F^2 + \|b\|_2^2 + \sum_{i=1}^{N} \|Z_i\|_F^2. \tag{25}$$

Structural zeros are excluded. Coordinates share the reactant-centered frame, and each comparison fixes atom identities, N, and padding. Expectations may average over these quantities when the stated constants hold uniformly. We use noiseless forward maps throughout; the reconstruction risk in the analysis is not the noise-regularized training loss.

E.1 Stability Derived from the Codec Architecture

Architectural conditions and domains. Fix finite parameters, atom metadata, and decoder mode. Forward evaluation is deterministic, LayerNorm divides by $\sqrt{\mathrm{var}(x) + \epsilon_{\mathrm{LN}}}$ with $\epsilon_{\mathrm{LN}} > 0$, and each valid attention query has at least one valid key. These conditions concern the stated architecture, not perfect reconstruction or successful transport. We evaluate latent maps on a closed ball $K_r = \{z : \|z\|_z \le r\}$. Encoder domains contain the coordinate segments being compared, including predicted completions when used in Theorem 1.

Raw dot-product attention need not be globally Lipschitz on an unbounded domain (Kim et al., 2021); bounded-domain analysis instead retains the dependence on feature magnitudes (Castin et al., 2024). We derive the required local bounds for the actual Gram-feature and vector-synthesis maps. No global attention bound is presumed.

Lemma 1 (Local codec stability)**.** *Under the conditions above, E_ϕ, D_ψ^ρ, and $H_\omega(\cdot, C)$ are continuously differentiable in their coordinate or active-code inputs. Let $c_\rho(z)$ stack all decoder coefficients $\alpha_{k,i,\ell}$ and $c_H(z, C)$ stack all internal-image head coefficients $\beta_{k,i,\ell}$. On K_r, define*

$$A_f = \sup_{z \in K_r} \|c_f(z)\|_2, \qquad J_f = \sup_{z \in K_r} \|\mathrm{D}c_f(z)\|_{\mathrm{op}}, \qquad f \in \{\rho, H\}. \tag{26}$$

Here C is fixed for the head. These quantities are finite, and valid stability constants are

$$\begin{aligned} L_{D^\rho} &= \frac{1}{\sqrt{3N}} + \frac{A_\rho + r J_\rho}{4\sqrt{3NT}}, \\ L_H &= \frac{A_H + r_H J_H}{\sqrt{3N(T-2)}}, \qquad r_H = \sqrt{r^2 + \max_i (\|R_i\|_2^2 + \|P_i\|_2^2)}. \end{aligned} \tag{27}$$

The decoder output norm is all-frame RMS, and the head output norm is internal-frame RMS. On a compact convex coordinate domain $\mathcal{K}$, a valid encoder constant is

$$L_E = \sup_{W \in \mathcal{K}} \|\mathrm{D}E_\phi(W)\|_{\mathrm{all} \to z} = \sqrt{3NT} \sup_{W \in \mathcal{K}} \|J_E(W)\|_2 < \infty, \tag{28}$$

where J_E uses unnormalized Euclidean coordinates.

Proof. The encoder uses fixed temporal projections, Gram features, SiLU MLPs, vector-coefficient products, and averaging over $N > 0$ valid atoms. These operations are smooth. The head uses the same operations with fixed endpoint vectors. Decoder attention adds affine maps, softmax, GELU, and additive-ϵ LayerNorm, which are smooth on valid queries. Neither tangent estimation nor peak selection enters these forward maps. Consequently, each coefficient map and its Jacobian are continuous; their suprema on K_r are finite. The same compactness argument and integration along coordinate segments give Equation 28.

For a stacked vector basis V_i, let B_V act on coefficients by $(B_V c)_{k,i} = V_i^\top c_{k,i}$. Its blocks repeat across frames, so $\|B_V\|_{\mathrm{op}} = \max_i \|V_i\|_{\mathrm{op}}$. For $V_i = [Z_i; G]$, this norm is at most r and

$$\|B_V(z) - B_V(z')\|_{\mathrm{op}} \le \|z - z'\|_z. \tag{29}$$

Adding and subtracting $B_V(z')c_\rho(z)$ therefore yields

$$\|B_V(z)c_\rho(z) - B_V(z')c_\rho(z')\|_2 \le (A_\rho + rJ_\rho)\|z - z'\|_z. \tag{30}$$

The temporal synthesis term has operator norm $\sqrt{T}$ because $\Phi^\top\Phi = TI_K$. Multiplying the correction by $1/4$ and converting the output norm by $1/\sqrt{3NT}$ gives L_{D^ρ}.

For the head, $V_i^H = [Z_i; G; R_i; P_i]$ has norm at most r_H, while its variation with z still has operator norm at most one. The head has neither a temporal synthesis term nor a $1/4$ multiplier. The same product bound, normalized by $\sqrt{3N(T-2)}$, gives L_H. □

Coefficient-network bounds. These constants can be bounded from the learned weights, rather than postulated independently. For $\Gamma_s(V) = \mathrm{vec}(VV^\top)/s$,

$$\mathrm{D}\Gamma_s(V)[\Delta V] = \mathrm{vec}(\Delta V V^\top + V\Delta V^\top)/s, \quad \|\mathrm{D}\Gamma_s(V)[\Delta V]\|_2 \le \frac{2\|V\|_F}{s}\|\Delta V\|_F. \tag{31}$$

The head uses $f_i^H = [h_i, b, \Gamma_1(V_i^H)]$ as input to a two-layer SiLU coefficient MLP with matrices W_1, W_2; h_i is the fixed atom-type embedding. Since $|\,\mathrm{SiLU}'\,| \le 1 + 1/e$, a valid bound is

$$J_H \le \|W_2\|_2(1 + 1/e)\|W_1\|_2\sqrt{N}\max\{1, 2r_H\}. \tag{32}$$

The decoder input is $d_{k,i} = [h_i, \Gamma_{25}([Z_i; G]), b, \eta(s_k), \rho]$, where η contains fixed frame features. Input and output projections $W_{\mathrm{in}}, W_\alpha$ and a Transformer stability bound $L_\mathcal{T}$ give

$$J_\rho \le \|W_\alpha\|_2 L_\mathcal{T}\|W_{\mathrm{in}}\|_2\sqrt{TN}\max\{1, 2r/25\}. \tag{33}$$

The factors account for stacking all coefficient inputs, including repeated shared features. Also $A_f \le \|c_f(0)\|_2 + rJ_f$. Mode ρ is fixed; shared weights do not imply equal constants or equal prediction risks in the two modes.

For completeness, additive-ϵ LayerNorm satisfies $\|\mathrm{D\,LN}\|_2 \le \|\gamma\|_\infty/\sqrt{\epsilon_{\mathrm{LN}}}$. For one attention head, write $S = \mathrm{softmax}_{\mathrm{row}}(Q_a K_a^\top/\sqrt{d_h})$ and $O = SV_a$. The softmax Jacobian has norm at most $1/2$, giving

$$\|\Delta O\|_F \le \|S\|_2\|\Delta V_a\|_F + \frac{\|V_a\|_2}{2\sqrt{d_h}}\left(\|K_a\|_2\|\Delta Q_a\|_F + \|Q_a\|_2\|\Delta K_a\|_F\right) \tag{34}$$

for differentials. On the relevant compact feature domain, the affine query/key/value maps, including biases, are bounded and $\|S\|_2 \le \sqrt{N}$. Composition with output projections, feedforward maps, and residual additions supplies a finite $L_\mathcal{T}$ for the finite stack of pre-norm layers.

What the regularity result does not assume. For a finite collection of finite-valued predictions, compact domains containing all comparison segments exist. A population-level uniform constant requires corresponding domain control; finite second moments alone do not provide it. The bounds establish finiteness, not smallness, and sampled Jacobian maxima are not certified upper bounds. Gaussian reconstruction noise has unbounded support, so no fixed code ball is claimed to cover every noisy training input. Theorems 1 and 3 concern noiseless maps and specify their own comparison domains.

E.2 Search-Response Separation and Retained Information

Theorem 2 (Search-response separation)**.** *For two eligible candidates with the same condition C, let $z_j = E_\phi(Y_j)$, $u_j = \mathcal{U}(Y_j)/\sigma_U$ for a fixed scale $\sigma_U > 0$, and $e_j = \|H_\omega(z_j, C) - u_j\|_{\text{int}}$. On a code ball containing z_1, z_2, let L_H be the head stability bound derived in Lemma 1. Then*

$$L_H \|z_1 - z_2\|_z \geq \big[\|u_1 - u_2\|_{\text{int}} - e_1 - e_2\big]_+, \tag{35}$$

where $[r]_+ = \max(r, 0)$.

Proof of Theorem 2. Write $h_j = H_\omega(z_j, C)$. Since the condition is identical for both candidates, the triangle inequality and Lemma 1 give

$$\|u_1 - u_2\|_{\text{int}} \leq \|u_1 - h_1\|_{\text{int}} + \|h_1 - h_2\|_{\text{int}} + \|h_2 - u_2\|_{\text{int}} \tag{36}$$

$$\leq e_1 + L_H \|z_1 - z_2\|_z + e_2. \tag{37}$$

Rearranging and taking the positive part proves the claim. Fixing C is essential: otherwise, the readout could explain the response difference through endpoint conditions alone. □

A direct connection to the search objective. There is also a distributional statement without a Lipschitz assumption. Let $Z = E_\phi(Y)$, $u = \mathcal{U}(Y)/\sigma_U$, and $\mu(Z, C) = \mathbb{E}[u \mid Z, C]$. Assume $\mathbb{E}\|u\|^2_{\text{int}} < \infty$ and $\mathcal{L}_{\text{search}} < \infty$. Conditional orthogonality yields

$$\begin{aligned} \mathcal{L}_{\text{search}} &= \mathbb{E}\|u - \mu(Z, C)\|^2_{\text{int}} + \mathbb{E}\|H_\omega(Z, C) - \mu(Z, C)\|^2_{\text{int}}, \\ \mathbb{E}\|u - \mu(Z, C)\|^2_{\text{int}} &\leq \mathcal{L}_{\text{search}}. \end{aligned} \tag{38}$$

To verify this identity, expand $H_\omega(Z, C) - u = (H_\omega(Z, C) - \mu) + (\mu - u)$. The cross term has zero conditional expectation given (Z, C). For variable atom counts, the argument is applied conditional on C, which fixes N and the coordinate normalization, and then averaged over conditions. Thus small search risk limits the response information left unresolved by the code and endpoints together. It need not make the code itself easy to predict from endpoints.

Physical scope. For a unit tangent, $I - 2t_k t_k^\top$ is an orthogonal reflection: climbing preserves force norm while reversing its tangential component. The other-image operator $I - t_k t_k^\top$ retains only the transverse component. These identities explain Equation 3; they establish neither an index-one saddle nor search convergence. Latent rescaling also changes L_H, so the separation bound is not a cross-tokenizer transportability measure.

E.3 From Completion to Endpoint-Conditioned Latent Prediction

Proof of Theorem 1. Let A be the event that every internal frame is masked and $q(X) = \Pr(A \mid X)$. On A, the retained boundaries are exactly $u = 0$ and $v = T - 1$; hence Equation 4 gives $\widetilde{X} = L$, $\mathcal{M} = \text{int}$, and $\mathcal{V} = \{0, T - 1\}$. Set $r_{\text{end}}(X) = \ell_{\text{int}}(D^1_\psi(E_\phi(L)), X)$ and $\mathcal{R}_{\text{end}} = \mathbb{E}_X r_{\text{end}}(X)$. Nonnegativity gives

$$\mathcal{L}_{\text{comp}} = \tfrac{1}{2}\mathbb{E}_{X,\mathcal{M}}\left[\ell_{\mathcal{M}}(\widehat{X}^{\,c}, X) + \ell_{\mathcal{V}}(\widehat{X}^{\,c}, X)\right] \tag{39}$$

$$\geq \tfrac{1}{2}\mathbb{E}_{X,\mathcal{M}}\left[\mathbf{1}_A\, \ell_{\text{int}}\big(D^1_\psi(E_\phi(L)), X\big)\right] \tag{40}$$

$$= \tfrac{1}{2}\mathbb{E}_X[q(X) r_{\text{end}}(X)] \geq \tfrac{p_{\min}}{2}\mathcal{R}_{\text{end}}. \tag{41}$$

This step requires only conditional coverage $q(X) \geq p_{\min}$, not independent masking. The stated uniform-length sampler supplies $q(X) = 1/(T - 2)$ because the unique span of maximal length masks all internal frames.

The endpoint-clamped completion $Q(C)$ and X share endpoints, so

$$\|Q(C) - X\|^2_{\text{all}} = \frac{T - 2}{T} r_{\text{end}}(X). \tag{42}$$

The encoder bound between these paths now gives

$$\mathbb{E}\|g(C) - E_\phi(X)\|^2_z \leq L_E^2 \mathbb{E}\|Q(C) - X\|^2_{\text{all}} \tag{43}$$

$$= L_E^2 \frac{T - 2}{T}\mathcal{R}_{\text{end}} \leq \frac{2L_E^2(T - 2)}{T p_{\min}}\mathcal{L}_{\text{comp}}. \tag{44}$$

This proves Equation 6. Both Q and g depend only on C and fixed models; no target internal coordinates enter their evaluation. The same bound also holds for displacement prediction, since subtracting the source code $z_s(C)$ from both $g(C)$ and $E_\phi(X)$ leaves their difference unchanged. □

Domains and a local refinement. The encoder domain must contain $[Q(C), X]$, including completion outputs outside the training support. No physical validity of $Q(C)$ is required. A more local statement uses

$$\kappa_E(X) = \int_0^1 \|\mathrm{D}E_\phi(Q(C) + t(X - Q(C)))\|_{\mathrm{all}\to z}\, dt. \tag{45}$$

If $q(X) > 0$ and $B_E = \operatorname{ess\,sup}_X \kappa_E(X)^2/q(X) < \infty$, the same proof gives

$$\mathbb{E}\|g(C) - E_\phi(X)\|_z^2 \le 2\frac{T-2}{T} B_E \mathcal{L}_{\mathrm{comp}}. \tag{46}$$

The main statement uses the transparent sufficient bound $B_E \le L_E^2/p_{\min}$. This exhibits both sources of looseness: weak coverage of endpoint-only inputs and amplification of coordinate errors by the encoder. The completion risk is mask-averaged; an arbitrary finite minibatch need not contain an endpoint-only example.

Conditional ambiguity. Fix N and define $x = \mathrm{vec}(X_{\mathrm{int}})/\sqrt{3N(T-2)}$. Let $f(C)$ be the identically normalized internal output of $D_\psi^1(E_\phi(L))$, a deterministic function of C. If $\mathbb{E}\|x\|_2^2 < \infty$ and the prediction risk is finite,

$$\mathbb{E}\|f(C) - x\|_2^2 = \mathbb{E}\|x - \mathbb{E}[x \mid C]\|_2^2 + \mathbb{E}\|f(C) - \mathbb{E}[x \mid C]\|_2^2. \tag{47}$$

Averaging over N gives the decomposition of $\mathcal{R}_{\mathrm{end}}$. Conditional variance captures endpoint ambiguity and can be zero for unique empirical endpoint pairs. The theorem therefore constructs a particular predictor with a risk bound; it does not assume a unique physical channel, or claim that $E_\phi(L)$ is already close to $E_\phi(X)$. Completion mode is used only to construct g. No equality between the two decoder modes, nor a capacity guarantee for the independently trained Flow, is needed.

E.4 From Latent Predictability to Decoded Generation

The separation of reconstruction and latent-model error also appears in distributional analyses of scientific latent models (Wang et al., 2025). Here we use paired target codes and the actual numerical output; the completion objective supplies an additional endpoint-conditioned reference predictor.

Theorem 3 (Completion-to-generation error bound)**.** *Under Theorem 1, let $\widehat{z}(C)$ be the actual numerical Flow output and $\delta_F = (\mathbb{E}\|\widehat{z}(C) - g(C)\|_z^2)^{1/2}$. Suppose the reconstruction decoder has a common stability constant L_D on the code segments between $\widehat{z}(C)$ and $E_\phi(X)$, with finite δ_F and clean reconstruction risk. Then*

$$\sqrt{\mathcal{R}_{\mathrm{gen}}} \le \sqrt{\mathcal{R}_{\mathrm{rec,clean}}} + L_D\varepsilon_c + L_D\delta_F, \tag{48}$$

where $\mathcal{R}_{\mathrm{gen}} = \mathbb{E}\|P_C D_\psi^0(\widehat{z}) - X\|_{\mathrm{all}}^2$ and $\mathcal{R}_{\mathrm{rec,clean}} = \mathbb{E}\|D_\psi^0(E_\phi(X)) - X\|_{\mathrm{all}}^2$.

Proof of Theorem 3. Write $z_t = E_\phi(X)$ and $\widehat{z} = \widehat{z}(C)$. Since X already has endpoints (R, P), replacing the endpoints of any W can only decrease its all-frame squared error:

$$\|P_C W - X\|_{\mathrm{all}}^2 = \frac{1}{3NT}\sum_{k=1}^{T-2}\sum_{i=1}^{N}\|W_{k,i} - X_{k,i}\|_2^2 \le \|W - X\|_{\mathrm{all}}^2. \tag{49}$$

Applying this inequality, the triangle inequality, and the decoder bound gives

$$\begin{aligned}
\|P_C D_\psi^0(\widehat{z}) - X\|_{\mathrm{all}} &\le \|D_\psi^0(\widehat{z}) - X\|_{\mathrm{all}} && (50)\\
&\le \|D_\psi^0(\widehat{z}) - D_\psi^0(z_t)\|_{\mathrm{all}} + \|D_\psi^0(z_t) - X\|_{\mathrm{all}} && (51)\\
&\le L_D\|\widehat{z} - z_t\|_z + \|D_\psi^0(z_t) - X\|_{\mathrm{all}}. && (52)
\end{aligned}$$

Restoring the same translation to both paths does not change the error. Taking L^2 norms over reactions and using Minkowski's inequality gives

$$\sqrt{\mathcal{R}_{\text{gen}}} \leq \sqrt{\mathcal{R}_{\text{rec,clean}}} + L_D \sqrt{\mathcal{R}_z}, \tag{53}$$

where $\mathcal{R}_z = \mathbb{E}\|\widehat{z} - E_\phi(X)\|_z^2$. A second application of Minkowski's inequality, now inserting the analytical predictor $g(C)$, yields

$$\sqrt{\mathcal{R}_z} \leq \left(\mathbb{E}\|\widehat{z} - g(C)\|_z^2\right)^{1/2} + \left(\mathbb{E}\|g(C) - E_\phi(X)\|_z^2\right)^{1/2} \leq \delta_F + \varepsilon_c. \tag{54}$$

Substitution proves Equation 48. No exact ODE solution is used: $\widehat{z}$ is the terminal output of the actual numerical solver. □

Local decoder sensitivity. A common L_D is sufficient, but not necessary for a pointwise statement. For $\gamma(t) = z_t + t(\widehat{z} - z_t)$, define

$$\kappa_D = \int_0^1 \|\mathrm{D}D_\psi^0(\gamma(t))\|_{z \to \text{all}}\, dt. \tag{55}$$

The same proof gives $\sqrt{\mathcal{R}_{\text{gen}}} \leq \sqrt{\mathcal{R}_{\text{rec,clean}}} + (\mathbb{E}[\kappa_D^2 \|\widehat{z} - z_t\|_z^2])^{1/2}$ whenever these quantities are finite. This weighted term cannot generally be factored into separate expectations. Lemma 1 supplies a common L_D on a shared bounded code domain, but does not establish that such a domain covers an arbitrary population.

Interpretation and unresolved approximation error. The reconstruction term concerns noiseless clean-target codes, not $\mathcal{L}_{\text{rec}}$, which mixes input types and adds latent noise. The term δ_F compares the actual Flow output with a concrete endpoint-conditioned predictor and can be evaluated from endpoints and fixed models. The implemented Flow is not trained to imitate g, and the theorem neither assumes zero discrepancy nor bounds it by $\mathcal{L}_{\text{FM}}$. Completion and re-encoding are used only in the proof, not added to generation. A collapsed encoder can make latent prediction trivial, but generally leaves reconstruction error; retaining both terms makes this degeneracy explicit. Comparisons between tokenizers must also account for their different stability constants. These results support the two pretraining requirements without proving improved fixed-capacity optimization, generalization, or physical channel validity.