DUAL FLOW MATCHING: LEARNING COUPLED FLOWS FROM PAIRED VIEWS OF A SHARED TARGET

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard flow matching predicts velocity from a single noisy state, leaving the relationship between paths toward the same data target implicit. We introduce *Dual Flow Matching* (DualFM), which couples two independently corrupted views of a shared target within one backbone. The network observes both states and predicts both velocities in a single evaluation. Independent training clocks vary the information available in each view, while synchronized sampling evolves the two states jointly. Training remains simulation-free and requires no pretrained teacher. We prove that auxiliary conditioning weakly reduces irreducible regression risk and derive an exact common-task identity for comparing field-approximation errors. Our evaluation spans MNIST, CIFAR-10, ImageNet32, and ImageNet64, including class-conditional generation with SiT-XL/4 at the larger resolution. Our experiments show improved multi-step generation over I-CFM, with approximately 0.1% additional parameters and nearly unchanged large-batch sampling latency. Matched controls show gains beyond model capacity and observation information, while diagnostics show more accurate target estimates and learned use of instance-level correspondence. The gains also extend to minibatch optimal transport coupling. These results establish shared-target joint prediction as an effective way to improve flow matching while keeping the backbone size nearly unchanged.

1 INTRODUCTION

Flow matching (FM) learns a generative vector field by regressing velocities along noise-to-data interpolations (Lipman et al., 2023; Liu et al., 2023). Different noise sources define different paths toward the same data example. Their intermediate states are linked by a shared endpoint, yet a standard single-state predictor observes only one noisy view at a time. We ask: *can paths toward the same target inform each other's velocity predictions?*

We introduce *Dual Flow Matching* (DualFM): two independent noise sources, two training clocks, and one shared target. For the same augmented image Y, construct

$$\boxed{X_t = (1-t)Y + t\epsilon_X, \qquad H_s = (1-s)Y + s\epsilon_H.} \tag{1}$$

Independent noise creates complementary views of the same target. One backbone reads both states and predicts both velocities in a single evaluation. Training samples t, s independently to vary the information in each view. Sampling synchronizes the clocks, jointly evolves the states, and returns only X. The construction remains simulation-free and needs neither a teacher nor a second backbone.

Figure 1 illustrates the auxiliary state's role: changing H at fixed X changes the continuation of the joint flow. Statistically, conditioning on H cannot increase the Bayes regression risk, with strict improvement whenever it changes the conditional mean. Both ideal branch velocities correspond to the same posterior target estimate. DualFM thus integrates complementary observations rather than ensembling independent output predictions.

Our analysis separates the target-bearing and difference coordinates and derives a common-task risk identity for source-matched comparisons. We evaluate DualFM on MNIST, CIFAR-10, Ima-

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.

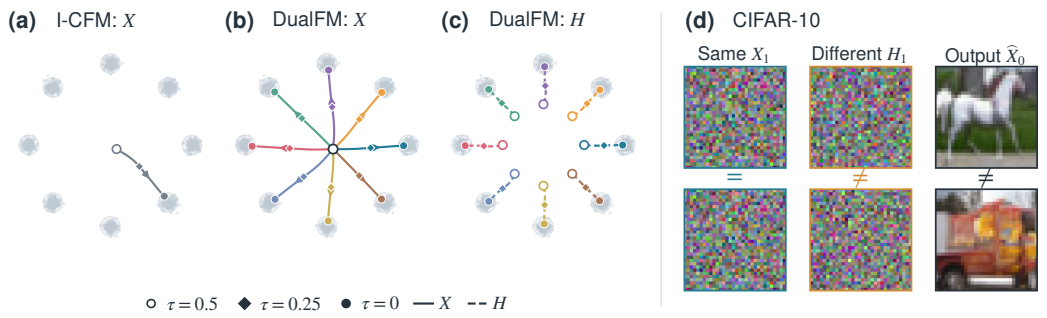

Figure 1: **The same sample state follows different paths with different auxiliary states.** *(a–c)* Learned toy trajectories from fixed $X_{0.5}$, with one I-CFM continuation or eight DualFM auxiliary states. Colors pair X (solid) and H (dashed); gray clouds show target modes. *(d)* Identical X_1 and different H_1 yield different CIFAR-10 images at Euler-100. Expanded views: Appendix D.

geNet32, and ImageNet64, covering UNets and the SiT-XL/4 transformer. ImageNet64 adds class-conditional generation at a larger resolution, while MNIST provides a small-scale validation. On CIFAR-10 and ImageNet32, DualFM lowers mean FID at all tested 20–1000 NFE budgets with nearly unchanged backbone cost. CIFAR-10 controls test capacity, source scale, interaction, and target correspondence. A matched minibatch OT experiment further shows that the generation gains extend beyond independent source–target pairing. Held-out diagnostics show approximately 20% lower reconstruction error than ordinary I-CFM. They also show sensitivity to auxiliary-target replacement and close agreement between the target estimates from the two branches. Our contributions are a single-backbone method for shared-target joint prediction, an exact analysis of its observations and regression objective, and controlled experiments on generation quality, cost, and target correspondence.

2 Related Work

Flow matching learns continuous transports through simulation-free velocity regression (Lipman et al., 2023; 2024); stochastic interpolants provide a broader framework for deterministic and stochastic dynamics (Albergo et al., 2025). Rectified flow refines source–target coupling through reflow (Liu et al., 2023), while minibatch coupling methods change how noise and data are paired (Pooladian et al., 2023; Tong et al., 2024). These methods operate on source–target assignments. DualFM instead couples two corrupted views of the assigned target. We use independent noise–data pairing in the main benchmarks and extend the construction to minibatch OT in Section 6.4.

AuxPath-FM adds an auxiliary random variable to the interpolation path (Peng & Gao, 2026). DualFM instead evolves a second image state from independent noise, supervises both states toward the same data instance, and predicts both velocities in one shared-backbone evaluation.

Representation-based methods offer another way to improve flow learning. REPA aligns features with a pretrained encoder (Yu et al., 2025), while REG jointly generates images and semantic tokens (Wu et al., 2025). Most closely, DUPA independently noises the same image at two clocks, regresses a velocity for each view, and aligns their intermediate condition features (Peng et al., 2026). Its velocity predictions process the views separately; their cross-view link is a training-time feature-alignment loss. DualFM instead conditions each velocity on *both* full image states, uses only velocity regression, and jointly evolves the states at inference. Self-Flow distills features across unequal noise levels (Chefer et al., 2026), whereas DualFM directly supervises two shared-target velocities.

UniDiffuser jointly predicts noise for paired modalities at separate timesteps in one network (Bao et al., 2023). DualFM instead constructs two views of the same image and supervises both toward that shared instance. PiFM studies path independence in multi-parameter flow matching (Téllez et al., 2026). DualFM uses two clocks to vary the information available in each view during training,

then follows a fixed synchronized path at inference. Its objective is shared-target learning along two coupled states, rather than transport consistency across arbitrary clock paths.

3 PRELIMINARIES

Let $Y \sim p_{\text{data}}$ and $\epsilon \sim \mathcal{N}(0, I_d)$ be independent, with $\mathbb{E}\|Y\|_2^2 < \infty$. Ordinary I-CFM uses

$$X_t = (1-t)Y + t\epsilon, \qquad U = \dot{X}_t = \epsilon - Y, \quad t \in [0,1], \tag{2}$$

and minimizes $\mathbb{E}\,\|v_\theta(X_t, t) - U\|_d^2$, where $\|a\|_d^2 = \|a\|_2^2/d$ and $t \sim \mathcal{U}(0,1)$. This is independent conditional flow matching (I-CFM) with no additional path noise (Tong et al., 2024); FM denotes the broader framework. Throughout, $t = 0$ denotes data and $t = 1$ noise; generation integrates from 1 to 0. The population minimizer $v^*(x,t) = \mathbb{E}[U \mid X_t = x]$ transports the prescribed marginals under suitable well-posedness conditions (Lipman et al., 2023). Individual ODE trajectories need not reproduce the sampled straight interpolations, which justifies simulation-free training.

4 DUAL FLOW MATCHING

Dual Flow Matching (DualFM) jointly evolves a sample state $X \in \mathbb{R}^d$ and an auxiliary state $H \in \mathbb{R}^d$ with one shared backbone. Figure 2 summarizes shared-target training and joint sampling.

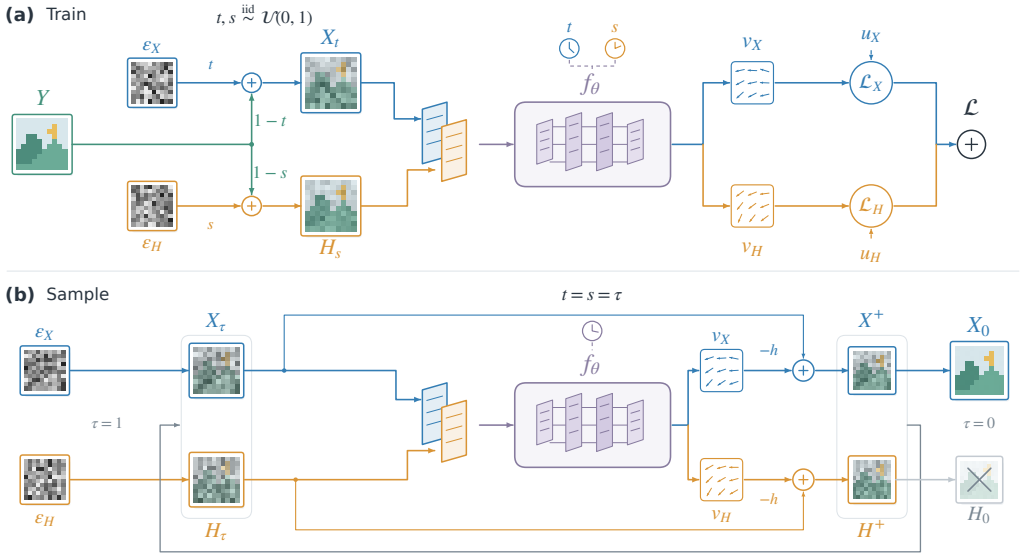

Figure 2: **Dual Flow Matching with one shared backbone.** (a) Independent sources and clocks form two views of Y; one backbone predicts both velocities. (b) Synchronized backward Euler updates evolve both states from the same pre-update pair; sampling returns only X. Image tiles and velocity symbols are schematic.

4.1 SHARED-TARGET PATHS AND SIMULATION-FREE TRAINING

Let $Y \sim p_{\text{data}}$, where p_{data} denotes the training-target distribution after preprocessing and augmentation, with $\mathbb{E}\|Y\|_2^2 < \infty$. Independently draw $\epsilon_X, \epsilon_H \sim \mathcal{N}(0, I_d)$ and $t, s \sim \mathcal{U}(0,1)$, and construct

$$X_t = (1-t)Y + t\epsilon_X, \qquad H_s = (1-s)Y + s\epsilon_H. \tag{3}$$

The two branches share the *same augmented data example* Y. At fixed clocks, the corruptions are independent given Y. The observations are linked through their shared target, which motivates joint prediction. Their native-time velocities are

$$U_X = \epsilon_X - Y, \qquad U_H = \epsilon_H - Y. \tag{4}$$

Writing $\|u\|_d^2 = \|u\|_2^2/d$, we train the joint predictor $(v_{\theta,X}, v_{\theta,H}) = f_\theta(X_t, H_s, t, s)$ with

$$\mathcal{L}_{\mathrm{DualFM}}(\theta) = \mathbb{E}\left[\|v_{\theta,X} - U_X\|_d^2 + \|v_{\theta,H} - U_H\|_d^2\right]. \tag{5}$$

Both branches use constant unit loss weights and clean endpoints: $X_0 = H_0 = Y$ in the prescribed interpolation. Equation (3) directly provides training states and targets, without simulation. Independent clocks let either view be more informative, encouraging cross-view prediction. This may help handle unequal state errors during synchronized sampling. The clock ablation in Section 6.3 establishes the empirical benefit of independent-clock training.

At the population optimum, squared regression gives the conditional-mean velocities. In particular, for $t, s > 0$, define $\mu(x, h, t, s) = \mathbb{E}[Y \mid X_t = x, H_s = h, t, s]$. Then

$$v_X^*(x, h, t, s) = \frac{x - \mu(x, h, t, s)}{t}, \qquad v_H^*(x, h, t, s) = \frac{h - \mu(x, h, t, s)}{s}. \tag{6}$$

Both ideal velocities use the same posterior target estimate μ, with motion determined by each branch's current state. The network directly predicts a concatenated velocity tensor, split into the two branch outputs. Section 6.5 evaluates learned target correspondence and agreement between the branch-implied clean estimates.

4.2 A SINGLE BACKBONE FOR TWO VELOCITIES

For image data, we concatenate X and H along channels and predict both velocities in the same channel layout. The UNet's internal feature widths, depth, and attention structure are unchanged from the single-state I-CFM baseline. Sinusoidal embeddings of t and s are concatenated and linearly projected to the input width of the original time-embedding MLP. Only the input/output interfaces and clock projection expand. Both tasks share the expensive feature hierarchy, which is evaluated once per call. For the RGB UNet, the parameter increase is approximately 0.1%. One network function evaluation (NFE) denotes one call returning *both* velocities. Table 3 summarizes the overhead; Table 11 provides parameter counts, arithmetic, memory, and sampling timings.

4.3 SYNCHRONIZED JOINT SAMPLING

At inference, we initialize $(\widehat{X}_1, \widehat{H}_1) = (\epsilon_X, \epsilon_H)$ with independent standard Gaussians and fix the clocks to $t = s = \tau$. The learned global-time ODE is

$$\frac{d}{d\tau}\begin{pmatrix}\widehat{X}_\tau \\ \widehat{H}_\tau\end{pmatrix} = F_\theta(\widehat{X}_\tau, \widehat{H}_\tau, \tau) := \begin{pmatrix}v_{\theta,X}(\widehat{X}_\tau, \widehat{H}_\tau, \tau, \tau) \\ v_{\theta,H}(\widehat{X}_\tau, \widehat{H}_\tau, \tau, \tau)\end{pmatrix}. \tag{7}$$

Because both inference clocks have derivative one, native and global velocities coincide here. Uniform Euler sampling uses $\tau_n = 1 - n/N$ and $\widehat{Z}_{\tau_{n+1}} = \widehat{Z}_{\tau_n} - F_\theta(\widehat{Z}_{\tau_n}, \tau_n)/N$, where $\widehat{Z} = (\widehat{X}, \widehat{H})$ and $n = 0, \ldots, N-1$. It requires exactly N NFEs and never evaluates the field at $\tau = 0$. The generated image is $\widehat{X}_0$. Sampling is deterministic conditional on the two initial sources. During training, H_s is a corrupted observation of the known target; during sampling, it evolves jointly with X from noise. No target image or external auxiliary observation is supplied at inference.

The associated ideal path is $Z_\tau = ((1-\tau)Y + \tau\epsilon_X, (1-\tau)Y + \tau\epsilon_H)$. Its conditional-mean joint velocity is the restriction of Equation (6) to the diagonal clocks. Under the well-posedness assumptions stated in Appendix A.1, the canonical population field transports these joint marginals for $\tau > 0$, and the marginal law converges to $\mathrm{Law}(Y, Y)$ as $\tau \downarrow 0$ (Lipman et al., 2023). The appendix explains the diagonal restriction and endpoint limit. Section 6.5 evaluates the learned diagonal field.

5 ANALYSIS OF SHARED-TARGET COUPLING

We analyze the regression risk, information in the observations, and joint objective of shared-target coupling. Proofs and the independent-clock extension appear in Appendix A.

5.1 AUXILIARY OBSERVATIONS AND CONDITIONAL REGRESSION RISK

Fix $t, s > 0$, let $m_0 = \mathbb{E}[U_X \mid X_t]$ and $m_1 = \mathbb{E}[U_X \mid X_t, H_s]$, and define their irreducible risks $\mathcal{B}_i = \mathbb{E}\,\|U_X - m_i\|_d^2$. Conditional expectation gives

$$\mathcal{B}_0 - \mathcal{B}_1 = \mathbb{E}\,\|m_1 - m_0\|_d^2 \geq 0. \tag{8}$$

Conditioning on H_s weakly reduces the Bayes regression risk. The reduction is strict if and only if $m_1 \neq m_0$ with positive probability. Recent FM methods explicitly model multimodal velocity distributions (Guo & Schwing, 2025; Chen et al., 2025). DualFM retains squared velocity regression and uses the auxiliary view to reduce target uncertainty. For any square-integrable predictor v measurable with respect to a conditioning set $\mathcal{C}$,

$$\mathbb{E}\,\|v - U_X\|_d^2 = \mathbb{E}\,\|v - \mathbb{E}[U_X \mid \mathcal{C}]\|_d^2 + \mathbb{E}\,\|U_X - \mathbb{E}[U_X \mid \mathcal{C}]\|_d^2\,. \tag{9}$$

Ordinary I-CFM and DualFM generally have different irreducible risks. To isolate field approximation, we compare DualFM with source-matched I-CFM on a common task, where the shared Bayes floor cancels exactly (Appendix A.4).

5.2 A COMMON GENERATIVE TASK FOR COMPARING LEARNED FIELDS

For the synchronized path, let $M = (X+H)/2$ and $D = (X-H)/2$. The sources $\bar{\epsilon} = (\epsilon_X + \epsilon_H)/2$ and $\epsilon_D = (\epsilon_X - \epsilon_H)/2$ are independent $\mathcal{N}(0, I_d/2)$ variables, giving

$$M_\tau = (1-\tau)Y + \tau\bar{\epsilon}, \qquad D_\tau = \tau\epsilon_D. \tag{10}$$

In particular, D_τ is independent of $(M_\tau, Y, \bar{\epsilon})$.

Proposition 1 (Common–difference decomposition)**.** *For $0 < \tau \leq 1$, the ideal synchronized joint field satisfies*

$$F_M^*(M, D, \tau) = b(M, \tau) := \mathbb{E}[\bar{\epsilon} - Y \mid M_\tau = M], \qquad F_D^*(M, D, \tau) = \frac{D}{\tau}, \tag{11}$$

where $F_M^* = (v_X^* + v_H^*)/2$ and $F_D^* = (v_X^* - v_H^*)/2$.

On the synchronized path, the common coordinate carries the target information and the ideal difference coordinate contracts analytically. Under independent training clocks, the branch-velocity difference also depends on the posterior target estimate: $v_X^* - v_H^* = x/t - h/s - (1/t - 1/s)\mu$. The target-dependent term cancels when $t = s$, but remains off the diagonal; thus, independent-clock training makes both velocity combinations depend on target estimation. The common source covariance $I_d/2$ defines a *source-matched I-CFM* control with the same target field $b(M, \tau)$. This provides a common task for comparing how the single-state and joint-state parameterizations learn the generative field.

5.3 COMPLEMENTARY OBSERVATIONS, CONSISTENT PREDICTIONS

The shared-target loss separates exactly into target-estimation and branch-consistency terms. For $t, s > 0$, define the clean estimates from the two branch velocities:

$$\widehat{Y}_X = X_t - t v_{\theta,X}, \qquad \widehat{Y}_H = H_s - s v_{\theta,H}. \tag{12}$$

Let $a = t^{-2}$, $b = s^{-2}$, and $\widehat{Y}_{\mathrm{av}} = (a\widehat{Y}_X + b\widehat{Y}_H)/(a+b)$. The samplewise loss in Equation (5) satisfies

$$\boxed{\ell_{\mathrm{DualFM}} = \underbrace{(a+b)\left\|\widehat{Y}_{\mathrm{av}} - Y\right\|_d^2}_{\text{target estimation}} + \underbrace{\frac{ab}{a+b}\left\|\widehat{Y}_X - \widehat{Y}_H\right\|_d^2}_{\text{cross-branch consistency}}.} \tag{13}$$

This identity holds for arbitrary network outputs (proof in Appendix A.6) and separates target estimation error from disagreement between the two clean estimates. Cross-view interaction comes from joint conditioning in the shared backbone; the decomposition adds no coupling loss. Sampling returns X alone. Section 6.5 tests the corresponding prediction structure through reconstruction, target replacement, and branch agreement.

To separate observation information from the learned parameterization, information-matched I-CFM compresses the paired observations into a sufficient statistic and learns a single predictor, with the induced time distribution, loss weight, and sampling clock (Appendix A.5). DualFM instead retains both views and learns two interacting velocity outputs under the shared-target objective. The inputs preserve the same posterior information, but the models represent and learn the velocity field differently. Table 4 tests this distinction: DualFM achieves better generation quality even after information matching. Appendix A.4 further derives an exact excess-risk identity for comparing fields on a common source-matched task.

6 Experiments

6.1 Experimental setup

CIFAR-10 and ImageNet32 are the main benchmarks, with MNIST as small-scale validation. Unconditional models use 600k updates, matched backbone sizes, RGB target batch 128, and EMA weights. The baseline is I-CFM (Equation (2)); ImageNet64 extends the comparison to class-conditional SiT-XL/4 at 1.2M updates. We report FID (Heusel et al., 2017) on 50,000 generated images using CleanFID (Parmar et al., 2022), with uniform Euler sampling at nine budgets from 2 to 1000 NFEs. Each DualFM NFE returns both velocities. For CIFAR-10 and ImageNet32, we report mean and sample SD over three training seeds. Protocols and per-run results are in Appendix B.

On MNIST (LeCun et al., 1998), DualFM improves on I-CFM at every reported budget (Table 1); the complete sweep is in Appendix B.5.

Table 1: **Generation quality on MNIST.** FID (↓) is computed from 50,000 samples generated with uniform Euler integration. Both methods use one training seed and 600k EMA weights; the lower value at each sampling budget is bold.

Method / NFE	10	20	50	100
I-CFM	4.427	2.554	1.880	1.725
DualFM	**4.213**	**2.348**	**1.836**	**1.631**

6.2 Generation quality and cost

Table 2: **Generation quality across datasets and backbones.** FID-50k from EMA checkpoints. (a) Unconditional UNets: 600k updates, mean ± sample SD over three runs. (b) Class-conditional SiT-XL/4: 1.2M updates, one training seed. Bold: lower mean. Complete UNet sweeps: Table 9.

Model / NFE	10	20	50	100	1000
(a) UNet 600k updates three training seeds					
CIFAR-10					
I-CFM	15.008 ±0.256	8.694 ±0.100	5.619 ±0.031	4.624 ±0.025	3.793 ±0.044
DualFM	**14.770** ±0.221	**7.797** ±0.113	**5.003** ±0.149	**4.290** ±0.132	**3.743** ±0.086
ImageNet32					
I-CFM	19.655 ±0.389	13.314 ±0.166	10.697 ±0.054	9.738 ±0.027	8.692 ±0.032
DualFM	**19.442** ±0.026	**12.509** ±0.090	**9.882** ±0.105	**9.113** ±0.119	**8.448** ±0.114
(b) SiT-XL/4 1.2M updates one training seed					
ImageNet64 (class-conditional)					
I-CFM	11.985	8.276	6.591	6.035	5.558
DualFM	**11.537**	**7.424**	**6.283**	**5.838**	**5.344**

DualFM lowers mean FID at eight of nine budgets on CIFAR-10 (Krizhevsky, 2009) and all nine on ImageNet32 (Deng et al., 2009; Chrabaszcz et al., 2017) (Tables 2 and 9). CIFAR-10 Euler-5 is the sole exception; ImageNet32 Euler-5 is nearly tied. The advantage holds throughout training: DualFM has lower Euler-50 FID at every measured checkpoint on both datasets (Figure 3).

With pixel-space SiT-XL/4 (Ma et al., 2024) on class-conditional ImageNet64, DualFM also achieves lower FID at every reported budget (Table 2(b)). Both methods use matched transformer

depth and width, shared class conditioning, and 1.2M updates. This extends the gains beyond UNets to a large transformer with four times as many image dimensions; implementation details are in Appendix B.1.

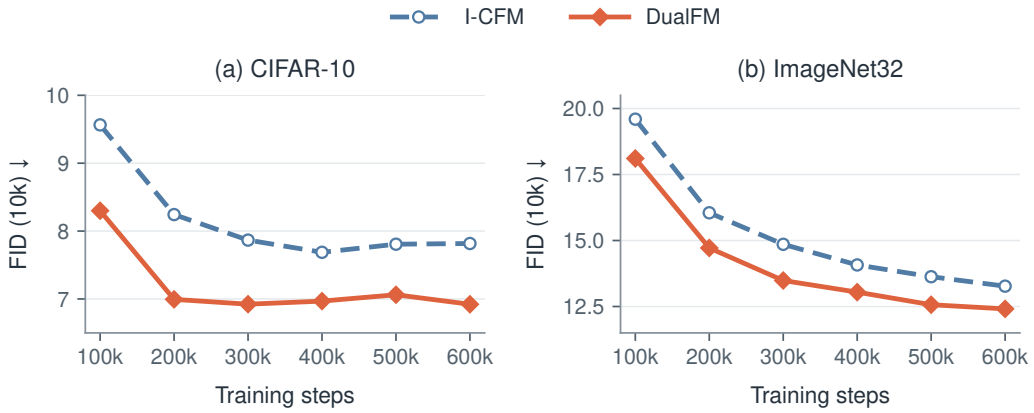

Figure 3: **DualFM's generation advantage persists throughout training.** Euler-50 FID on CIFAR-10 (left) and ImageNet32 (right): 10,000 EMA samples per checkpoint, one training run per curve.

On the UNet benchmarks, these gains come with only 0.104% additional parameters and 0.121% additional profiled forward arithmetic (Table 3). At batch 256, Euler-50 latency differs by at most 0.004 seconds in the two measured timing blocks. Thus the generation gains require almost no additional backbone computation.

Table 3: **Negligible backbone overhead.** Relative to I-CFM; H100, FP32, Euler-50. Detailed timings and memory: Table 11.

Parameters	Forward FLOPs	Batch-256 latency
+0.104%	+0.121%	Nearly identical ($\leq$ 0.004 s difference)

6.3 What makes coupling effective?

Table 4: **CIFAR-10 controls.** Mean FID-50k over three training seeds after 600k updates. Shared-target uses dedicated ablation runs; all DualFM variants sample synchronously with clean endpoints. Bold: best FID. Per-seed results: Table 10.

Method	Euler20	Euler50	Euler100	Euler1000
I-CFM	8.771	5.642	4.636	3.769
Parameter-matched I-CFM	8.793	5.683	4.671	3.803
Source-matched I-CFM	8.481	5.355	4.515	3.843
Information-matched I-CFM	8.308	5.321	4.509	3.817
No interaction (state budget)	8.507	5.479	4.512	3.850
No interaction (target budget)	9.209	5.915	4.974	4.224
DualFM, same-class targets	15.301	11.960	10.819	9.673
DualFM, random targets	21.426	17.666	16.282	14.805
DualFM, shared target (sync clocks)	8.552	5.533	4.767	4.136
DualFM, shared target	**8.091**	**5.172**	**4.417**	**3.693**

We compare capacity, observation, and coupling controls under a matched 600k-update training budget (Table 4). Parameter-matched I-CFM adds an active adapter; source-matched I-CFM uses noise standard deviation $1/\sqrt{2}$ (Proposition 1). No-interaction controls process views separately with shared weights, matching either the number of single-state inputs or the number of distinct

targets. Matching targets requires more computation. DualFM outperforms all four controls and ordinary I-CFM at every reported budget.

To control for the information available across the full training distribution, information-matched I-CFM uses a sufficient statistic with matched time distribution, loss weight, and sampling clock (Appendix A.5). DualFM outperforms this control across all four sampling budgets, reducing FID by 2.0–3.2%. The coupled model improves generation even when the observation information and induced training and sampling protocols are matched.

We next isolate clock sampling and target correspondence within the dual architecture. The synchronized-clock control sets $t = s = \tau$ during training, retaining independent noises and the same Shared-Y architecture, loss, and budget. Both variants sample synchronously. Independent-clock training reduces FID by 5.4–10.7% across all four budgets. Thus, training views with unequal information levels improves synchronized generation on CIFAR-10. To test target correspondence with architecture and budgets fixed, H targets the same image, another same-class image, or a random image. Each branch learns its assigned endpoint. Shared-Y outperforms both alternatives at every reported budget; same-class pairing improves over random pairing but underperforms ordinary I-CFM. Pairing by class therefore does not replace pairing by the exact same instance.

6.4 Beyond independent source–target pairing

Minibatch OT-CFM changes noise–data pairing; DualFM changes joint prediction. We combine them on CIFAR-10 using the exact OT solver of Tong et al. (2024). *Dual-OT* matches each Gaussian source pair to one image using the sum of the two branch transport costs. Both models share the backbone width, optimizer, global OT pool size, and 600k-update budget.

Dual-OT achieves lower FID than OT-CFM at every reported sampling budget (Table 5). Shared-target joint prediction therefore remains beneficial when minibatch transport replaces independent source–target pairing, extending the gains beyond the I-CFM setting.

Table 5: **Shared-target prediction with minibatch OT coupling.** CIFAR-10 FID-50k at 600k updates, using EMA weights and uniform Euler sampling. Both models use the same training and evaluation protocol; each Dual-OT NFE predicts both branch velocities.

Method	Euler10	Euler20	Euler50	Euler100	Euler1000
OT-CFM	13.417	9.048	6.112	5.059	4.081
Dual-OT	**12.086**	**7.413**	**5.010**	**4.289**	**3.736**

6.5 Integrating shared-target observations

We measure target reconstruction, sensitivity to target correspondence, and agreement between branches on held-out interpolation states with matched image and noise draws (Appendix B.9). For $\widehat{Y}_X = X_t - tv_{\theta,X}$, DualFM reduces $\mathbb{E}\left\|\widehat{Y}_X - Y\right\|_d^2$ by 20.7% on CIFAR-10 and 20.2% on ImageNet32, improving at all five times (Table 12 in Appendix C). Figure 4(a,b) fixes the primary noise time t within each column and varies the auxiliary time s; smaller s means a cleaner auxiliary view. Gains grow with auxiliary information. The dashed diagonal, extracted in panel (c), uses the synchronized clocks of sampling. Gains remain positive at all evaluated diagonal times on both datasets, so the benefit is not confined to asymmetric noise levels.

Beyond reconstruction, the common-task comparison isolates field approximation. Against the same target $U_M = \bar{\epsilon} - Y$, DualFM's common field has lower risk than source-matched I-CFM at all five evaluated CIFAR-10 times (Figure 5). The shared Bayes floor cancels by Equation (22). The lower risk therefore shows better common-field approximation, beyond the benefit of additional observation information.

The reconstruction benefit depends on the relationship between the views. Unlike the target-pairing ablations in Section 6.3, these interventions keep the trained model fixed. The ablations ask which pairing helps the model learn; the interventions ask whether the resulting predictor uses the correspondence between its inputs. Replacing only the target used to construct H, while fixing all other inputs and the X target, raises CIFAR-10 reconstruction MSE from 0.0422 to 0.1279 for random

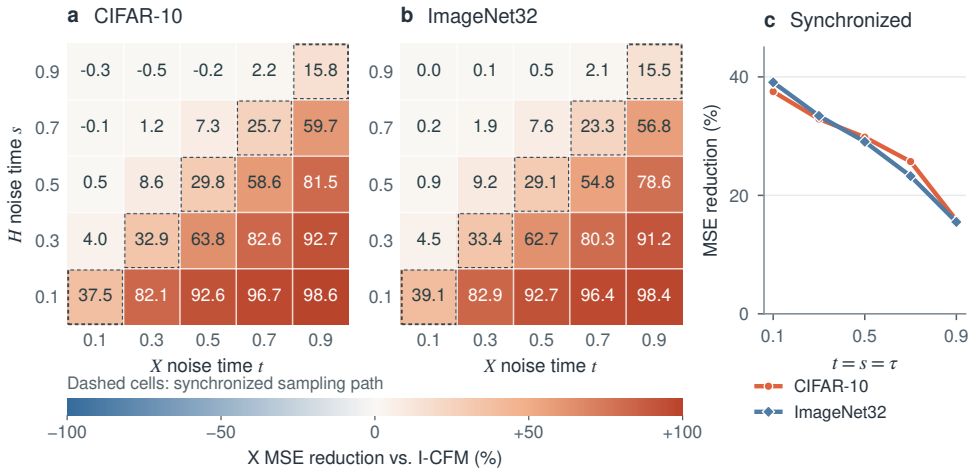

Figure 4: **Auxiliary observations improve target estimation across noise levels.** X reconstruction MSE reduction versus I-CFM at matched t, over 1,024 held-out images per dataset. Dashed cells in (a,b) mark synchronized clocks, plotted in (c); negative values indicate higher MSE.

replacements (Table 12). Same-class replacements also increase error (Appendix C). The dedicated Shared-Y model shows the same sensitivity, whereas the model trained on random targets is nearly unaffected. This selective sensitivity demonstrates learned use of instance-level correspondence.

The cosine similarity between branch residuals exceeds 0.999 on both datasets, and averaging gives negligible benefit. The no-interaction control has cosine 0.568 and benefits from fusion. This supports a common target estimate (Equation (6)), rather than errors that cancel when outputs are averaged. Sampling returns X alone. The two-spiral probe in Appendix C.4 further separates reduced mean error from averaging-induced variance reduction. Actual generation shows the same branch agreement: on ImageNet32, X and H FID differ by less than 0.046 across five sampling budgets, with per-coordinate endpoint MSE below 9×10^{-6}. Averaging the endpoints does not improve X's FID (Appendix C.5). Unlike interpolation-based diagnostics, these measurements test branch agreement along actual sampling trajectories, with both states evolved from Gaussian noise.

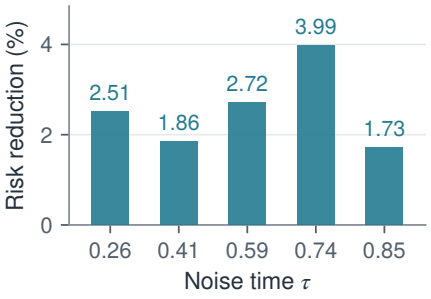

Figure 5: **Lower common-task risk.** DualFM's target MSE reduction over source-matched I-CFM on CIFAR-10. Full risks: Appendix C.1.

7 Conclusion

DualFM makes shared-target structure explicit in flow matching: one backbone jointly predicts velocities for two independently corrupted views of the same data instance. Training remains simulation-free. Experiments show improved multi-step generation over I-CFM on MNIST, CIFAR-10, and ImageNet32, as well as class-conditional ImageNet64 with SiT-XL/4. On CIFAR-10 and ImageNet32, these gains come with only 0.104% additional parameters and nearly unchanged large-batch sampling latency. Our analysis shows that auxiliary observations weakly reduce irreducible regression risk and provides a common task for comparing learned fields. Matched controls and diagnostics support the role of shared-target coupling, with more accurate target estimates and lower common-task field risk. The gains also persist with minibatch OT coupling.

Future work will explore shared-target coupling at higher resolutions and in other modalities.

Reproducibility Statement

Section 4 specifies the probability paths, regression targets, network interfaces, and sampler. Appendix A supplies the assumptions and proofs, including the derivations of the source- and information-matched controls evaluated in Table 4. Appendix B details preprocessing, training, evaluation, interventions, and statistical units, with checkpoint identities, training and generation seeds, rank counts, and sample budgets recorded for reproducibility. Anonymous code for data preparation, training, and evaluation of the CIFAR-10 and ImageNet32 DualFM main experiments is available at `https://anonymous.4open.science/r/DualFM-412C/`.

Ethics Statement

This work concerns generative modeling on existing image benchmarks and does not involve a new human-subject study. Generative models can reproduce dataset biases or memorized content and can be misused to produce misleading imagery. Broader use should respect the source datasets' access conditions and assess privacy, fairness, and misuse risks in the deployment context.

AI Use Statement

In this work, we used generative AI tools only to aid or polish manuscript writing, for retrieval and discovery (including finding related work), and for research ideation or execution. No generative AI tools were used for other purposes. The authors retain responsibility for the final content and its validation.

References

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL `https://www.jmlr.org/papers/v26/23-1605.html`.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 1692–1717. PMLR, 2023. URL `https://proceedings.mlr.press/v202/bao23a.html`.

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

Hansheng Chen, Kai Zhang, Hao Tan, Zexiang Xu, Fujun Luan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. Gaussian mixture flow matching models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9783–9802, 2025. URL `https://proceedings.mlr.press/v267/chen25cl.html`.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of ImageNet as an alternative to the CIFAR datasets. *arXiv preprint arXiv:1707.08819*, 2017. URL `https://arxiv.org/abs/1707.08819`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. URL `https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf`.

Pengsheng Guo and Alex Schwing. Variational rectified flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20921–20940, 2025. URL `https://proceedings.mlr.press/v267/guo25i.html`.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

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.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024. URL https://arxiv.org/abs/2412.06264.

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.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2401.08740.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022. URL https://arxiv.org/abs/2104.11222.

Changhao Peng, Yuqi Ye, Shuangjun Du, Wenxu Gao, and Wei Gao. Dual-path condition alignment for diffusion transformers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e8bd617e7dd0394ceadf37b4a7773179-Abstract-Conference.html.

Xin Peng and Ang Gao. Flow matching with arbitrary auxiliary paths. *arXiv preprint arXiv:2605.06364*, 2026. URL https://arxiv.org/abs/2605.06364.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127, 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Francisco Téllez, AmirHossein Zamani, Philippe Martin, Shuang Ni, Guy Wolf, Eugene Belilovsky, Sina Sanjari, and Yanlei Zhang. Path-independent flow matching for multi-parameter generative dynamics. *arXiv preprint arXiv:2605.13487*, 2026. URL https://arxiv.org/abs/2605.13487.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier than you think. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/0b99315234cc95e6ef281f9155b68832-Abstract-Conference.html.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.06940.

A MATHEMATICAL DETAILS

All squared vector norms below are normalized per coordinate, $\|u\|_d^2 = d^{-1}\|u\|_2^2$. We assume $\mathbb{E}\|Y\|_2^2 < \infty$ and the independent Gaussian sources specified in Section 4. Statements about regression apply to square-integrable predictors. Clocks are held fixed unless an integration over them is explicitly stated.

A.1 JOINT TRANSPORT AND BOUNDARY BEHAVIOR

For $t, s > 0$, the interpolation equations imply

$$U_X = \frac{X_t - Y}{t}, \qquad U_H = \frac{H_s - Y}{s}. \tag{14}$$

Taking conditional expectations gives Equation (6). At synchronized clocks, set $p_\tau = \mathrm{Law}(Z_\tau)$ and $U = (\epsilon_X - Y, \epsilon_H - Y)$, so that $\dot{Z}_\tau = U$ for the prescribed interpolation process. For a smooth compactly supported test function $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$,

$$\frac{d}{d\tau}\mathbb{E}[\varphi(Z_\tau)] = \mathbb{E}[\nabla\varphi(Z_\tau)^\mathsf{T} U] \tag{15}$$

$$= \mathbb{E}[\nabla\varphi(Z_\tau)^\mathsf{T}\mathbb{E}[U \mid Z_\tau]] = \mathbb{E}[\nabla\varphi(Z_\tau)^\mathsf{T} F^*(Z_\tau, \tau)]. \tag{16}$$

This is the weak continuity equation for p_τ and the canonical conditional-mean field F^*. On each interval $[T, 1]$ with $T > 0$, assume that this field has a unique nonexplosive ODE flow. Also assume that the corresponding continuity equation has a unique backward-time solution with the prescribed noise law at $\tau = 1$. The ODE marginals then equal p_τ on that interval.

Two-clock training versus diagonal evaluation. The transport result uses the canonical conditional-mean field on the synchronized path. A continuous representative, when it exists, is unique on the interior support. Bounding the learned field's error on the diagonal requires additional regularity. The two-clock risk identifies the field almost everywhere, and even continuous errors can concentrate near the diagonal. Synchronized field diagnostics and the clock-training ablation are reported in Section 6.5 and Table 4, respectively.

The endpoint is a distributional limit. The interpolation satisfies

$$\mathbb{E}\|Z_\tau - (Y, Y)\|_2^2 = \tau^2\mathbb{E}\big[\|\epsilon_X - Y\|_2^2 + \|\epsilon_H - Y\|_2^2\big] \longrightarrow 0. \tag{17}$$

Consequently p_τ converges to the law of (Y, Y) as $\tau \downarrow 0$. The limiting law is supported on the diagonal of $\mathbb{R}^{2d}$; transport at $\tau = 0$ is therefore understood through this distributional limit. The finite learned sampler returns its X branch without imposing equality between the two numerical endpoints.

A.2 CONDITIONAL RISK AND THE VALUE OF AN AUXILIARY OBSERVATION

For fixed clocks, write $m_0 = \mathbb{E}[U_X \mid X_t]$ and $m_1 = \mathbb{E}[U_X \mid X_t, H_s]$. The residual decomposition is $U_X - m_0 = (U_X - m_1) + (m_1 - m_0)$. Its cross term vanishes because $\mathbb{E}[U_X - m_1 \mid X_t, H_s] = 0$. Taking squared norms and expectations gives

$$\mathbb{E}\,\|U_X - m_0\|_d^2 = \mathbb{E}\,\|U_X - m_1\|_d^2 + \mathbb{E}\,\|m_1 - m_0\|_d^2, \tag{18}$$

which proves Equation (8). The same orthogonality argument for $v - U_X = (v - \mathbb{E}[U_X \mid \mathcal{C}]) + (\mathbb{E}[U_X \mid \mathcal{C}] - U_X)$ proves Equation (9). The reduction is nonnegative, and is strictly positive exactly when the conditional means differ with positive probability.

A.3 PROOF OF PROPOSITION 1

The vectors $\bar{\epsilon}$ and ϵ_D are jointly Gaussian, with

$$\mathrm{Cov}(\bar{\epsilon}) = \mathrm{Cov}(\epsilon_D) = I_d/2, \qquad \mathrm{Cov}(\bar{\epsilon}, \epsilon_D) = 0. \tag{19}$$

They are therefore independent, and both are independent of Y. The linear interpolation gives Equation (10) and hence independence of D_τ and (M_τ, U_M), where $U_M = \bar{\epsilon} - Y$. Their native joint-path targets transform as

$$\frac{U_X + U_H}{2} = U_M, \qquad \frac{U_X - U_H}{2} = \epsilon_D. \tag{20}$$

Conditioning on (M_τ, D_τ) yields $F_M^* = \mathbb{E}[U_M \mid M_\tau] = b(M_\tau, \tau)$ and $F_D^* = \epsilon_D = D_\tau/\tau$. This proves Proposition 1. In particular, the ideal difference trajectory satisfies $D_\tau = \tau D_1$, contracting as sampling time decreases. The common coordinate has source law $\mathcal{N}(0, I_d/2)$ and target field b, which define the source-matched comparison below.

For an arbitrary two-head predictor, define $F_{\theta,D} = (v_{\theta,X} - v_{\theta,H})/2$ and $F_{\theta,M} = (v_{\theta,X} + v_{\theta,H})/2$. The parallelogram identity gives the samplewise synchronized loss decomposition

$$\|v_{\theta,X} - U_X\|_d^2 + \|v_{\theta,H} - U_H\|_d^2 = 2\,\|F_{\theta,M} - U_M\|_d^2 + 2\,\|F_{\theta,D} - \epsilon_D\|_d^2\,. \tag{21}$$

This decomposition separates learning of the common generative field from fitting the analytic difference field, motivating the common-task evaluation.

A.4 Learning the common generative field

Use identical held-out draws from Equation (10) to compare $F_{\theta,M} = (v_{\theta,X} + v_{\theta,H})/2$ at $(X, H, t, s) = (M + D, M - D, \tau, \tau)$ with a source-matched prediction $u_\psi(M, \tau)$. Both predict $U_M = \bar{\epsilon} - Y$. All expectations use the same evaluation distribution.

Proposition 2 (Common-task excess-risk comparison)**.** *Let* $R_{\mathrm{D}} = \mathbb{E}\,\|F_{\theta,M} - U_M\|_d^2$ *and* $R_{\mathrm{S}} = \mathbb{E}\,\|u_\psi - U_M\|_d^2$ *at fixed* $\tau > 0$*. Then*

$$R_{\mathrm{D}} - R_{\mathrm{S}} = \mathbb{E}\,\|F_{\theta,M} - b\|_d^2 - \mathbb{E}\,\|u_\psi - b\|_d^2\,. \tag{22}$$

The shared Bayes floor cancels, leaving the difference in approximation errors. The identity allows the learned DualFM field to depend on D and can be integrated over a common time distribution when the risks are finite.

Proof and nuisance decomposition. Let $g = F_{\theta,M}(M_\tau, D_\tau, \tau)$, $f = u_\psi(M_\tau, \tau)$, and $b = b(M_\tau, \tau)$. Here b is the Bayes field for the evaluation distribution; it also describes the training distribution when the two distributions agree. Since D_τ is independent of (M_τ, U_M),

$$\mathbb{E}[U_M - b \mid M_\tau, D_\tau] = 0. \tag{23}$$

Both $g - b$ and $f - b$ are measurable with respect to (M_τ, D_τ), so their inner products with $U_M - b$ have expectation zero. Consequently

$$\mathbb{E}\,\|g - U_M\|_d^2 = \mathbb{E}\,\|g - b\|_d^2 + \mathbb{E}\,\|U_M - b\|_d^2\,, \tag{24}$$

$$\mathbb{E}\,\|f - U_M\|_d^2 = \mathbb{E}\,\|f - b\|_d^2 + \mathbb{E}\,\|U_M - b\|_d^2\,. \tag{25}$$

Subtracting proves Equation (22).

Next define $\bar{g} = \mathbb{E}[g \mid M_\tau]$, averaging over the correct conditional law $D_\tau \sim \mathcal{N}(0, \tau^2 I_d/2)$. Independence additionally gives $\mathbb{E}[g - \bar{g} \mid M_\tau, U_M] = 0$. Expanding $g - U_M = (g - \bar{g}) + (\bar{g} - U_M)$ therefore yields

$$\mathbb{E}\,\|g - U_M\|_d^2 = \mathbb{E}\,\|\bar{g} - U_M\|_d^2 + \mathbb{E}\,\|g - \bar{g}\|_d^2\,, \tag{26}$$

which separates prediction risk after averaging over D from variation with D. Resampling D at fixed M estimates this variation.

Finite-resampling estimator. For each fixed (M_τ, U_M), draw $K \geq 2$ independent difference states from their conditional Gaussian law and write the predictions $g_1, \ldots, g_K$, with $\bar{g}_K = K^{-1} \sum_k g_k$. The sample variance

$$\widehat{V}_D = \frac{1}{K-1} \sum_{k=1}^{K} \|g_k - \bar{g}_K\|_d^2 \tag{27}$$

is unbiased for the conditional difference-induced variance. To see this, expand $\sum_k \|g_k - \bar{g}_K\|_2^2 = \sum_k \|g_k - \bar{g}\|_2^2 - K\|\bar{g}_K - \bar{g}\|_2^2$; conditional independence gives variance $1/K$ for the sample mean. Averaging over held-out draws therefore estimates $V_D = \mathbb{E}\,\|g - \bar{g}\|_d^2$ without a finite-K downward bias. By the same independence,

$$\mathbb{E}\,\|\bar{g}_K - U_M\|_d^2 = \mathbb{E}\,\|\bar{g} - U_M\|_d^2 + \frac{V_D}{K}. \tag{28}$$

The finite-resample mean retains variance V_D/K. Uncertainty is computed over independent held-out examples, grouping repeated difference states and time points within each example.

A.5 Sufficiency, iso-information orbits, and the single-state control

Fix $t, s > 0$, set $a_X = (1-t)/t$, $a_H = (1-s)/s$, and write $A = X_t/t = a_X Y + \epsilon_X$ and $B = H_s/s = a_H Y + \epsilon_H$. For $r = \sqrt{a_X^2 + a_H^2} > 0$, the transformation

$$\begin{pmatrix} Q \\ R \end{pmatrix} = \frac{1}{r}\begin{pmatrix} a_X & a_H \\ -a_H & a_X \end{pmatrix}\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} rY + \epsilon_Q \\ \epsilon_R \end{pmatrix} \tag{29}$$

is orthogonal in the two noise coordinates. Thus ϵ_Q, ϵ_R are independent standard Gaussians and R is independent of (Y, Q) at fixed clocks. Because the observation transform is invertible, $\mathrm{Law}(Y \mid X_t, H_s, t, s) = \mathrm{Law}(Y \mid Q, r)$. The $r > 0$ construction covers all clock pairs except $t = s = 1$, a probability-zero corner with no target information.

Orbit construction. Fix a target draw Y, an independent $\epsilon_Q \sim \mathcal{N}(0, I_d)$, and $r > 0$, and set $Q = rY + \epsilon_Q$. For each $\phi \in [0, \pi/2]$, choose $a_X = r\cos\phi$, $a_H = r\sin\phi$, $t = 1/(1 + a_X)$, and $s = 1/(1 + a_H)$. With $R \sim \mathcal{N}(0, I_d)$ independent of (Y, Q), invert the transform:

$$X_t = \frac{\cos\phi\, Q - \sin\phi\, R}{1 + a_X}, \qquad H_s = \frac{\sin\phi\, Q + \cos\phi\, R}{1 + a_H}. \tag{30}$$

For each fixed angle these are valid original interpolation draws. Use the same independently drawn R across angles to form a paired orbit, and repeat with fresh independent R to estimate nuisance variation. Every angle shares the same posterior observation (Q, r). Each target draw is one statistical unit across angles and repetitions. The sufficient-statistic single-state estimator supplies a reference that is invariant to R and ϕ by construction.

Matching the full training information. Set $u = 1/(1 + r)$ and

$$W = \frac{Q}{1 + r} = (1 - u)Y + u\epsilon_Q, \qquad U_W = \epsilon_Q - Y. \tag{31}$$

For a common clean estimate $\widehat{Y}$, define native branch predictions $\widetilde{v}_X = (X_t - \widehat{Y})/t$, $\widetilde{v}_H = (H_s - \widehat{Y})/s$, and $f = Q - (1+r)\widehat{Y}$. Then $\widetilde{v}_X - U_X = -(1+a_X)(\widehat{Y} - Y)$, $\widetilde{v}_H - U_H = -(1+a_H)(\widehat{Y} - Y)$, and $f - U_W = -(1 + r)(\widehat{Y} - Y)$. Hence the samplewise identity is

$$\|\widetilde{v}_X - U_X\|_d^2 + \|\widetilde{v}_H - U_H\|_d^2 = \omega(t, s)\,\|f - U_W\|_d^2, \qquad \omega(t, s) = \frac{(1 + a_X)^2 + (1 + a_H)^2}{(1 + r)^2}. \tag{32}$$

For nonnegative finite clock odds, $1 \le \omega \le 2$. Indeed, writing $q = a_X + a_H$ gives $r \le q \le \sqrt{2}r$; the numerator minus $(1 + r)^2$ is $1 + 2(q - r) \ge 0$, while twice $(1 + r)^2$ minus the numerator is $r^2 + 4r - 2q \ge 0$. The information-matched I-CFM trains $f_\psi(W, u)$ by sampling the original independent clocks, constructing W, U_W, and applying this weight. It must inherit the induced distribution of u, which is not uniform. Given r, the remaining clock allocation is independent of (Y, ϵ_Q). The weight changes how much each effective time contributes to the loss, but leaves the conditional Bayes predictor of U_W given (W, u) unchanged. This sufficient-statistic single-state model provides the information-matched reference for the jointly conditioned two-head network.

Matching the physical sampling clock. Let $\sigma = 1/\sqrt{2}$ and $c_\tau = 1 - \tau + \sigma\tau$. The synchronous common path can be written $M_\tau = c_\tau W_{u_\tau}$, with $u_\tau = \sigma\tau/c_\tau$ and a unit-noise source for W. The corresponding common field induced by f_ψ is

$$F_{M,\psi}(M, \tau) = \frac{(\sigma - 1)M + \sigma f_\psi(M/c_\tau, \sigma\tau/c_\tau)}{c_\tau}. \tag{33}$$

To obtain this, use $c'_\tau = \sigma - 1$ and $u'_\tau = \sigma/c_\tau^2$ in the product and chain rules. The control begins at $M_1 \sim \mathcal{N}(0, I_d/2)$ and uses the same physical τ grid as DualFM, with one network call per step. The effective-time steps are induced by this physical grid. DualFM returns the learned X endpoint, whereas this control returns the common endpoint M, providing two finite-network realizations of the common generative task.

A.6 An exact loss decomposition for unconstrained velocity heads

Equation (32) uses a shared clean estimate. For DualFM's two separately predicted estimates, the following exact decomposition holds. Define

$$\widehat{Y}_X = X_t - t v_{\theta,X}, \qquad \widehat{Y}_H = H_s - s v_{\theta,H}, \qquad c_X = t^{-1}, \quad c_H = s^{-1}. \tag{34}$$

Let $K = c_X^2 + c_H^2$ and $\widehat{Y}_{\mathrm{av}} = (c_X^2 \widehat{Y}_X + c_H^2 \widehat{Y}_H)/K$. Write $e_i = \widehat{Y}_i - Y$ for $i \in \{X, H\}$ and $e_{\mathrm{av}} = \widehat{Y}_{\mathrm{av}} - Y$. The relation $c_X^2(e_X - e_{\mathrm{av}}) + c_H^2(e_H - e_{\mathrm{av}}) = 0$ cancels the cross term when expanding the weighted squared errors. The remaining centered term equals $(c_X^2 c_H^2/K)\,\|e_X - e_H\|_d^2$, yielding

$$\begin{aligned} &\|v_{\theta,X} - U_X\|_d^2 + \|v_{\theta,H} - U_H\|_d^2 \\ &\quad = K \left\| \widehat{Y}_{\mathrm{av}} - Y \right\|_d^2 + \frac{c_X^2 c_H^2}{K} \left\| \widehat{Y}_X - \widehat{Y}_H \right\|_d^2. \end{aligned} \tag{35}$$

If $f_{\mathrm{av}} = Q - (1 + r)\widehat{Y}_{\mathrm{av}}$, the first term is $\omega\,\|f_{\mathrm{av}} - U_W\|_d^2$. The second term measures disagreement between the two clean estimates. Both $\widehat{Y}_{\mathrm{av}}$ and f_{av} may depend on the nuisance coordinate and the clock allocation, unlike a predictor whose only inputs are (W, u). The identity separates weighted target-estimation error from cross-branch disagreement; interaction is supplied by the jointly conditioned backbone.

B EXPERIMENTAL AND IMPLEMENTATION DETAILS

The CIFAR-10 and ImageNet32 DualFM implementation, including data-preparation scripts, training configurations, and evaluation commands, is available at `https://anonymous.4open.science/r/DualFM-412C/`.

B.1 DATA, ARCHITECTURE, AND OPTIMIZATION

MNIST images are evaluated at 28×28 resolution. The MNIST I-CFM and DualFM models use the same hidden width and optimization protocol within that benchmark; DualFM changes only the state and clock interfaces required by the two-state construction. Both use 600k updates, training seed 42, EMA weights, and clean sample endpoints. DualFM samples its two training clocks independently and synchronizes them at inference, exactly as in the main method.

Our RGB experiments use CIFAR-10 and ImageNet32 at native 32×32 resolution. Pixel values are mapped from $[0, 255]$ to $[-1, 1]$. Training uses random horizontal flips with probability $1/2$ and no additional augmentation. CIFAR-10 uses a shuffled distributed data loader; ImageNet32 uses a memory-mapped image array with independently sampled training indices. Both branches receive the same *already augmented* target, except in the explicitly identified target-pairing controls. Class labels are never inputs to the generator. The same-class control uses labels only to construct its training pairs. All sampling is unconditional.

Table 6 gives the common optimization recipe; each control changes only the ingredients described below. Both DualFM endpoints are clean, and the auxiliary loss has unit weight throughout training.

Table 6: RGB UNet architecture and optimization configuration.

Setting	Value
Resolution and channels	32×32, RGB
Base feature width	128
Channel multipliers	$(1, 2, 2, 2)$
Residual blocks per level	2
Attention downsampling factor	2
Dropout	0.1
Optimizer	Adam (Kingma & Ba, 2015), $(\beta_1, \beta_2) = (0.9, 0.999)$
Adam ϵ; weight decay	10^{-8}; 0
Maximum learning rate	2×10^{-4}
Warm-up	5,000 linear steps
Gradient norm clipping	1
Updates	600,000
Global target batch	128, except the state-budget control
EMA decay	0.9999
Training arithmetic	FP32
DualFM clocks during training	Independent $\mathcal{U}(0, 1)$
DualFM source laws	Independent $\mathcal{N}(0, I)$
DualFM auxiliary loss weight	1 throughout training

The learning rate remains constant after warm-up. Checkpoints store the model, EMA, optimizer, and completed update count.

For ImageNet64, the pixel-space SiT-XL/4 has 28 blocks, width 1152, 16 attention heads, and 4×4 patches. Both models use 1000-class conditioning; DualFM expands the patch interfaces to six channels and combines two clock embeddings with one shared class embedding. Training uses seed 42, batch 128, FP32 with gradient checkpointing, Adam at 10^{-4}, 5000 warm-up updates, zero weight decay, gradient clipping at 1, and EMA decay 0.9999. The comparison uses 1.2M EMA weights, 50,000 samples, and ImageNet64 training-set FID statistics.

B.2 TRAINING AND SAMPLING ALGORITHMS

The algorithms specify the batching and update order in Section 4. Training targets are augmented once, before constructing either view; all Gaussian sources and clocks are independent across examples and branches. $\|\cdot\|_d^2$ averages image coordinates.

Algorithm 1: Shared-target DualFM training

1. Initialize network parameters θ and EMA parameters $\bar{\theta} = \theta$.
2. For each training update, draw a minibatch $\{Y_i\}_{i=1}^{B}$ and apply the dataset's augmentation once to each image.
3. Independently draw $\epsilon_{X,i}, \epsilon_{H,i} \sim \mathcal{N}(0, I)$ and $t_i, s_i \sim \mathcal{U}(0, 1)$.
4. Construct $X_i = (1 - t_i)Y_i + t_i\epsilon_{X,i}$ and $H_i = (1 - s_i)Y_i + s_i\epsilon_{H,i}$.
5. In one batched forward pass, compute $(v_{X,i}, v_{H,i}) = f_\theta(X_i, H_i, t_i, s_i)$.
6. Minimize $B^{-1}\sum_i[\|v_{X,i} - (\epsilon_{X,i} - Y_i)\|_d^2 + \|v_{H,i} - (\epsilon_{H,i} - Y_i)\|_d^2]$. Average gradients across training ranks, clip their global norm to 1, and take an Adam step with the prescribed warm-up schedule.
7. Update $\bar{\theta} \leftarrow 0.9999\bar{\theta} + 0.0001\theta$. Repeat from step 2; retain EMA weights for evaluation.

Algorithm 2: Synchronized Euler sampling (N NFEs)

1. Draw independent $X_1, H_1 \sim \mathcal{N}(0, I)$; set $h = 1/N$.
2. For $n = 0, \ldots, N - 1$, set $\tau_n = 1 - n/N$ and evaluate $(v_X^n, v_H^n) = f_{\bar{\theta}}(X_{\tau_n}, H_{\tau_n}, \tau_n, \tau_n)$ once.
3. Simultaneously set $(X_{\tau_{n+1}}, H_{\tau_{n+1}}) = (X_{\tau_n} - hv_X^n, H_{\tau_n} - hv_H^n)$ using the pre-update pair. Repeat steps 2–3 for all N time steps.
4. Return X_0 only. Convert from model range to image range for evaluation; do not average branch endpoints or inject trajectory noise.

B.3 Configuration differences between controls

Table 7 summarizes the ingredients changed in the CIFAR-10 ablations. "Targets" counts distinct augmented images per global update; "inputs" counts states, not shared-backbone evaluations. Two interacting states are concatenated along channels, whereas no-interaction states are concatenated along the batch dimension. All arms use clean endpoints, 600k updates, and EMA evaluation. Precise adapter and pairing implementations follow in Appendix B.7.

Table 7: **Configuration map for the CIFAR-10 controls.** Source standard deviations apply in both training and sampling. For two-state models, clocks are independent during training and synchronized at inference. Y' is another assigned endpoint, not added endpoint noise.

Model	Source std.	Interaction	H target	Targets	State inputs
Ordinary I-CFM	1	–	–	128	128
Parameter-matched I-CFM	1	–	–	128	128
Source-matched I-CFM	$1/\sqrt{2}$	–	–	128	128
No interaction, state budget	1,1	No	Y	64	128
No interaction, target budget	1,1	No	Y	128	256
DualFM, Shared-Y	1,1	Yes	Y	128	256
DualFM, same class	1,1	Yes	Y': same class	128	256
DualFM, random targets	1,1	Yes	Y': random	128	256

The parameter-matched and both no-interaction models retain the active adapter. Ordinary and source-matched I-CFM use the single-stream backbone; DualFM expands only input/output and clock interfaces. The three dedicated pairing runs use the same global-batch communication and permutation code, including the identity permutation for Shared-Y. The dedicated Shared-Y and main-benchmark DualFM models are separate training runs.

B.4 Sampling and evaluation

The Euler budget grid is $N \in \{2, 5, 10, 20, 50, 100, 200, 500, 1000\}$. All methods use uniform physical time $\tau_n = 1 - n/N$ and the sampling conventions in Algorithm 2: simultaneous branch updates, one NFE per joint call, and X-only output.

The FID protocol uses 50,000 generated images, the training-set reference, and CleanFID's `legacy_tensorflow` feature pipeline. Comparisons use the same reference statistics and Inception checkpoint. The common distributed evaluation recipe uses batch size 256 per rank and evaluation seed 2025, with rank seed offset $1{,}000{,}003$.

The main benchmark uses training seeds 42, 43, and 44. We compute the mean and sample standard deviation from the per-run scores, using denominator $n-1$ for the sample variance. These are run-level statistics, including training and sampling variation. For held-out diagnostic uncertainty, the resampling unit is a whole data example, with its clock points and perturbations grouped together.

Time convention. The implementation uses the reverse clock $t_{\mathrm{old}} = 1 - t$. Paper velocities therefore denote $v_{\mathrm{new}}(x, h, t, s) = -v_{\mathrm{old}}(x, h, 1-t, 1-s)$ (and analogously for I-CFM). This time change preserves generated states, losses, and FID. All plots and per-time tables use the paper's convention.

B.5 Complete MNIST Euler sweep

Table 8 expands Table 1 to all nine budgets under the protocol above. Both rows use training seed 42 and 600k EMA weights. The main text shows the mid-range interval $10 \le N \le 100$.

Table 8: **Complete MNIST Euler-budget evaluation.** FID (↓) is computed from 50,000 samples. Both methods use training seed 42 and 600k EMA weights; the lower value at each budget is bold.

Method / NFE	2	5	10	20	50	100	200	500	1000
I-CFM	57.774	10.768	4.427	2.554	1.880	1.725	1.587	1.420	1.371
DualFM	**54.682**	**9.757**	**4.213**	**2.348**	**1.836**	**1.631**	**1.496**	**1.354**	**1.221**

B.6 Complete CIFAR-10 and ImageNet32 results

Table 9 summarizes 108 evaluation records: two datasets, two models, three training seeds (42, 43, and 44), and nine Euler budgets. The accompanying records contain full-precision scores and checkpoint identities.

CIFAR-10 DualFM seed 44 resumed at 500k with model and optimizer states; the random generator and data-loader position were restarted.

Table 9: **Complete Euler-budget comparison.** FID-50k, 600k EMA checkpoints; mean ± sample SD over three training runs. Bold: lower mean within each dataset.

	CIFAR-10		ImageNet32	
NFE	I-CFM	DualFM	I-CFM	DualFM
2	170.758 ±1.449	**148.867** ±2.981	199.578 ±1.826	**180.078** ±5.231
5	**38.383** ±0.469	38.934 ±0.534	45.547 ±0.836	**45.542** ±0.928
10	15.008 ±0.256	**14.770** ±0.221	19.655 ±0.389	**19.442** ±0.026
20	8.694 ±0.100	**7.797** ±0.113	13.314 ±0.166	**12.509** ±0.090
50	5.619 ±0.031	**5.003** ±0.149	10.697 ±0.054	**9.882** ±0.105
100	4.624 ±0.025	**4.290** ±0.132	9.738 ±0.027	**9.113** ±0.119
200	4.162 ±0.045	**3.993** ±0.116	9.128 ±0.034	**8.722** ±0.120
500	3.873 ±0.031	**3.816** ±0.098	8.780 ±0.027	**8.509** ±0.116
1000	3.793 ±0.044	**3.743** ±0.086	8.692 ±0.032	**8.448** ±0.114

B.7 Capacity, source, and interaction controls

Table 4 reports mean FID over training seeds 42, 43, and 44 per method, using 600k EMA checkpoints, evaluation seed 2025, and eight evaluation ranks. Table 10 gives the individual results. The dedicated Shared-Y runs use the same global target-pairing protocol as the pairing controls; they are separate from the main-benchmark runs.

Table 10: Individual CIFAR-10 control results underlying Table 4. FID-50k after 600k updates using EMA weights; training seeds 42, 43, and 44, evaluation seed 2025, and eight evaluation ranks.

Method	Seed	Euler20	Euler50	Euler100	Euler1000
I-CFM	42	8.634	5.683	4.709	3.741
	43	8.823	5.556	4.617	3.805
	44	8.856	5.687	4.582	3.761
Parameter-matched I-CFM	42	8.857	5.651	4.689	3.852
	43	8.642	5.777	4.738	3.790
	44	8.880	5.621	4.586	3.767
Source-matched I-CFM	42	8.590	5.281	4.548	3.796
	43	8.437	5.376	4.454	3.857
	44	8.416	5.408	4.543	3.876
Information-matched I-CFM	42	8.252	5.404	4.480	3.839
	43	8.429	5.284	4.495	3.766
	44	8.243	5.275	4.552	3.846
No interaction (state budget)	42	8.544	5.418	4.564	3.832
	43	8.593	5.551	4.468	3.911
	44	8.384	5.468	4.504	3.807
No interaction (target budget)	42	9.038	5.971	4.907	4.255
	43	9.303	5.812	4.986	4.172
	44	9.286	5.962	5.029	4.245
DualFM, same-class targets	42	15.613	11.869	10.986	9.554
	43	15.117	12.188	10.578	9.716
	44	15.173	11.823	10.893	9.749
DualFM, random targets	42	21.139	17.864	16.128	14.896
	43	21.837	17.350	16.354	14.948
	44	21.302	17.784	16.364	14.571
DualFM, shared target (sync clocks)	42	8.680	5.484	4.849	4.072
	43	8.379	5.515	4.798	4.155
	44	8.597	5.600	4.654	4.181
DualFM, shared target	42	7.997	5.198	4.359	3.734
	43	8.128	5.253	4.491	3.664
	44	8.148	5.065	4.401	3.681

Parameter-matched I-CFM. This control remains a single-state I-CFM with a standard Gaussian source. It adds a residual adapter at the lowest-resolution UNet representation: two biased 1×1 convolutions with widths $256 \to 77 \to 256$, a SiLU nonlinearity, and a scalar-clock affine map to 27 channels, zero-padded to the adapter width. The up projection and clock map are zero-initialized. The adapter contributes exactly 39,811 trainable parameters, giving 38,347,014 parameters in total, equal to DualFM. The adapter is active during prediction.

Source-variance-matched I-CFM. For $a_0 = 1/\sqrt{2}$, this single-state control trains

$$W_\tau = (1-\tau)Y + \tau a_0 \epsilon, \qquad U_W = a_0\epsilon - Y, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{36}$$

Its sampler starts from $a_0\epsilon$ and integrates the learned velocity without an additional output rescaling, using the same source law in training and inference. This matches the synchronous mean path $M = (X + H)/2$ of DualFM, whose source covariance is $I/2$; the information-matched control below additionally matches the full independent-clock observation information.

Synchronized training clocks. This control retains the dedicated Shared-Y model's two-clock network, independent Gaussian sources, clean endpoints, summed branch loss, and global target-pairing communication. Only the training time sampler changes: instead of independent t, s, it uses $t = s = \tau$ with uniform τ. Batch size 128, optimizer, and the 600k-update budget remain fixed across the three-seed comparison. Both models use synchronous Euler inference and the same FID

protocol. Their mean FIDs at $N \in \{2, 5, 10, 20, 50, 100, 200, 500, 1000\}$ are respectively 166.287, 43.700, 16.179, 8.552, 5.533, 4.767, 4.392, 4.198, and 4.136.

Information-matched single-state I-CFM. Unlike source matching, this control matches the posterior information of the full independent-clock training distribution. It samples the original (t, s) and trains on the sufficient-statistic state W at effective time u, with weight $\omega(t, s)$ from Equation (32). Sampling uses the transformed common field in Equation (33) on the same physical time grid as DualFM, not a uniform effective-time grid. One network call is used per step, and the final common state is returned.

No-interaction controls and their budgets. The two noisy views are passed separately through a shared single-state network: $v_X = q_\theta(X, t)$ and $v_H = q_\theta(H, s)$. Stacking the views along the batch dimension allows one batched call, but each prediction uses only its own view. The parameter-matching adapter is retained. The unique-target-matched version uses the same 128 augmented targets per update as DualFM and evaluates 256 single-state inputs. It matches target exposure, while requiring extra backbone computation. The state-budget-matched version instead uses 128 single-state inputs constructed from 64 targets. It matches the number of backbone inputs but changes the number of unique targets. The two controls match target exposure and state-input budget, respectively. Each sampling branch is an ordinary single-state flow, and only the designated output branch is scored.

B.8 Target-pairing controls

All three pairing arms use the same global batch of already augmented CIFAR-10 images and the same communication operations. The X endpoint is Y_i, while the H endpoint is $Y_{\pi(i)}$. The shared-target arm sets $\pi(i) = i$. The same-class arm forms a randomly ordered cycle within each class; classes represented by a single item retain that item. The random-target arm forms a randomly ordered cycle over the whole batch. Thus each arm preserves the augmented-image batch through a finite-batch permutation.

Each branch's target velocity uses its assigned endpoint. The two clocks remain independent during training, both endpoints remain clean, and inference remains synchronous. A separate step-indexed random generator chooses permutations, leaving the noise, clock, and dropout streams unchanged. Labels determine the same-class pairs but are not generator inputs.

B.9 Held-out paired-view reconstruction diagnostics

Table 12 uses 1,024 images sampled without augmentation from the CIFAR-10 test split or ImageNet32 validation split, diagnostic seed 2025, and frozen 600k EMA weights. The main DualFM and ordinary I-CFM checkpoints use training seed 42. The auxiliary comparisons use the dedicated Shared-Y, random-target, and target-budget no-interaction checkpoints from the CIFAR-10 ablation. Network evaluation uses FP32 and batch 32 per GPU. Eight GPUs divide the time points, each of which uses all 1,024 images. Gaussian draws are indexed by example and shared across clocks and models. We average squared errors over image coordinates, then examples and the five synchronized times $0.1, 0.3, 0.5, 0.7, 0.9$ with equal weights. The full DualFM diagnostic also evaluates the corresponding 5×5 clock grid, but the reported table restricts to synchronized clocks to compare with ordinary I-CFM.

For the correspondence intervention, a fixed permutation replaces the target image used in H, keeping its Gaussian source and all other inputs fixed. The random permutation has no fixed points; the CIFAR-10 same-class permutation selects another image within the class. Permutation preserves the empirical auxiliary-target marginal while changing instance correspondence and the joint input distribution. Reconstruction is always scored against the original X target. Residual cosine is computed per example between $\widehat{Y}_X - Y$ and $\widehat{Y}_H - Y$, then averaged in the same way as MSE. Fusion uses the fixed weight $1/2$, without fitting weights. For no interaction, the same single-stream model is evaluated separately on each noisy view. Reported differences characterize these fixed checkpoints on held-out interpolation draws.

Supplemental replacement evaluations. Replacement evaluations retain the same 1,024 X examples, indexed noise, batch size 32, and frozen checkpoints. For ImageNet32, labels are checked against the original validation archive. Same-class H targets are assigned by a permutation with no fixed points within each class of the full 50,000-image validation pool. This ensures a different target for every selected X and preserves the full-pool target distribution. CIFAR-10 uses within-subset permutations.

Complete model comparisons, time-resolved curves, and per-clock intervention records appear in Appendix C.

Training-dynamics evaluation. The training-dynamics curves in Figure 3 use training seed 42 and generation seed 2025, with 10,000 samples from each EMA checkpoint and Euler-50 sampling. Lines connect the six measured checkpoints without smoothing.

B.10 Efficiency protocol

Parameter counts include every inference parameter. We estimate forward FLOPs with `torch.profiler`, counting a multiply–add as two operations. The estimate excludes unsupported operations and backward computation. Table 11 reports parameter counts and peak allocated memory separately.

Sampling-loop timings use one H100 80GB HBM3 per independent timing block, FP32, and no autocast, compilation, or distributed communication. Two blocks reverse model order, and each measurement takes the median of five repeats after one warm-up. The reference throughput setting is batch 256 with 50 Euler steps, with batch-one latency considered separately. Timed work includes clock-vector construction, neural evaluations, simultaneous state updates, and final device synchronization. Initial noise is already on device. Data loading, initial random draws, checkpoint I/O, image encoding, and FID computation are excluded. Peak memory is PyTorch's maximum allocated memory, including the resident baseline.

Table 11: **Model and sampling cost.** Single H100, FP32, Euler-50. Slashes separate timing blocks; latency uses the indicated batch size. GFLOPs count profiler-supported operators per image; peak memory uses batch 256. Protocol: Appendix B.

Dataset	Model	Params (M)	GFLOPs	Time_{256} (s) blocks 0 / 1	Time_{1} (s) blocks 0 / 1	Peak GiB blocks 0 / 1
CIFAR-10	I-CFM	38.307	11.774	4.848 / 4.825	0.528 / 0.459	3.094 / 3.125
	DualFM	38.347	11.788	4.849 / 4.828	0.456 / 0.465	3.131 / 3.131
ImageNet32	I-CFM	38.307	11.774	4.845 / 4.825	0.444 / 0.438	3.124 / 3.125
	DualFM	38.347	11.788	4.849 / 4.828	0.453 / 0.454	3.131 / 3.131

C COMMON-TASK AND PAIRED-VIEW EVIDENCE

C.1 COMMON-TASK FIELD APPROXIMATION ON CIFAR-10

We compare $g = (v_{\theta,X} + v_{\theta,H})/2$ and the source-matched field f against $U_M = \bar{\epsilon} - Y$ on the synchronized path. Figure 6 complements the main-text comparison in Figure 5 with the absolute risk difference. Each risk includes the same Bayes floor, so their difference estimates the difference in excess approximation risk in Equation (22). At $\tau \approx 0.261, 0.414, 0.586, 0.739, 0.850$, the risk differences are $-0.002283, -0.001517, -0.002374, -0.004316, -0.002399$, respectively.

For the DualFM measurements, the synchronized orbit has $\phi = 45°$ and $\tau = 1/(1 + r/\sqrt{2})$ for $r \in \{4, 2, 1, 0.5, 0.25\}$, in that order. The averaged clean estimate obeys $\widehat{Y}_M = M_\tau - \tau g$, so its squared reconstruction error divided by τ^2 gives exactly $\|g - U_M\|_d^2$. We score each of the four difference-state draws separately and report their mean loss at each evaluated time.

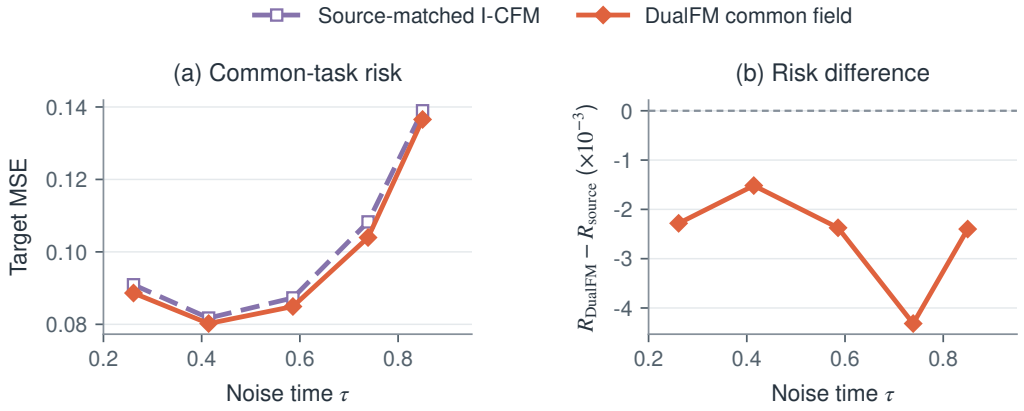

Figure 6: **Lower common-task field risk on CIFAR-10.** Left: per-coordinate velocity target MSE for the DualFM common field and source-matched I-CFM. Right: their risk difference, whose population counterpart equals the excess-approximation-risk difference. Negative values favor DualFM. Markers show the five evaluated times; lines connect their means.

C.2 COMPLETE RECONSTRUCTION AND CORRESPONDENCE RESULTS

Table 12 compares the main models and controls using the protocol in Appendix B.9. The correct-pair arm uses the same target for both observations, including for models trained with different endpoint pairings. The main DualFM and dedicated Shared-Y control are distinct training runs. Residual cosine and equal-fusion MSE use this correct-pair arm; all reconstruction errors are scored against the X target.

Figure 7 shows the results at each noise level; the main-text clock atlas additionally evaluates off-diagonal clocks. These interventions test sensitivity to target correspondence by changing the joint input distribution; the common-task comparison instead holds the prediction task fixed.

Table 12: **Complete synchronized correspondence and consistency comparison.** All completed diagnostic models; equal mean over five times and 1,024 images. Columns report X clean-target MSE under each auxiliary endpoint, residual cosine, and fixed equal-fusion MSE. Dashes denote inapplicable single-state I-CFM entries.

Model	Same image	Same class	Random	Cosine	Fusion
CIFAR-10					
I-CFM	0.053214	–	–	–	–
DualFM (main)	0.042217	0.120699	0.127871	0.999235	0.042213
Shared-Y control	0.042225	0.121700	0.128887	0.999277	0.042220
Same-class control	0.052721	0.053182	0.054046	0.559370	0.045138
Random-target control	0.053815	0.053686	0.053683	0.551529	0.045559
No interaction (target budget)	0.053478	0.053478	0.053478	0.567679	0.045550
ImageNet32					
I-CFM	0.059409	–	–	–	–
DualFM (main)	0.047403	0.131859	0.138850	0.999389	0.047401

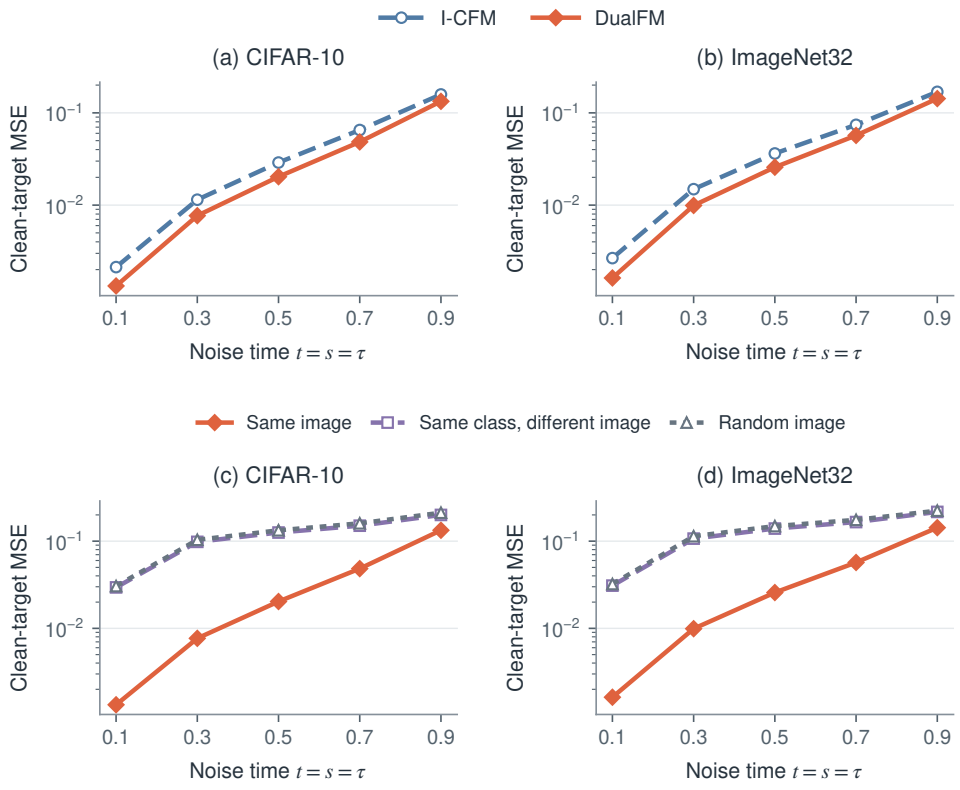

Figure 7: **Paired-view evidence across noise levels.** Top: reconstruction MSE for I-CFM and DualFM. Bottom: DualFM with the original or a replaced auxiliary target. Each point averages 1,024 held-out examples at a fixed checkpoint; vertical axes are logarithmic.

C.3 PER-TIME INTERVENTION RECORDS

Tables 13 and 14 give the individual clock measurements underlying Table 12 and Figure 7. The CIFAR-10 records include all six checkpoints, including the no-interaction target-budget control.

Table 13: **ImageNet32 per-time intervention results.** Each entry is X clean-target MSE over 1,024 held-out images; these are not FID scores.

Model	τ	Same image	Same class	Random
I-CFM	0.1	0.002664	–	–
	0.3	0.014891	–	–
	0.5	0.036295	–	–
	0.7	0.074198	–	–
	0.9	0.168998	–	–
DualFM (main)	0.1	0.001623	0.030913	0.032218
	0.3	0.009913	0.106958	0.113954
	0.5	0.025749	0.139446	0.148772
	0.7	0.056946	0.165550	0.175573
	0.9	0.142784	0.216431	0.223735

Table 14: **CIFAR-10 per-time intervention results.** Each entry is X clean-target MSE over 1,024 held-out images; these are not FID scores.

Model	τ	Same image	Same class	Random
I-CFM	0.1	0.002129	–	–
	0.3	0.011447	–	–
	0.5	0.028943	–	–
	0.7	0.065183	–	–
	0.9	0.158369	–	–
DualFM (main)	0.1	0.001330	0.029489	0.030533
	0.3	0.007683	0.098176	0.103607
	0.5	0.020313	0.125533	0.133149
	0.7	0.048426	0.150415	0.160133
	0.9	0.133334	0.199880	0.211931
Shared-Y control	0.1	0.001330	0.031731	0.032923
	0.3	0.007680	0.099612	0.105243
	0.5	0.020243	0.126696	0.134517
	0.7	0.048436	0.150802	0.160172
	0.9	0.133436	0.199660	0.211582
Same-class control	0.1	0.002209	0.002198	0.002210
	0.3	0.011897	0.011824	0.011994
	0.5	0.029715	0.029534	0.030144
	0.7	0.065128	0.065104	0.066643
	0.9	0.154657	0.157251	0.159237
Random-target control	0.1	0.002212	0.002200	0.002199
	0.3	0.011993	0.011911	0.011909
	0.5	0.030136	0.029942	0.029947
	0.7	0.066199	0.065924	0.065925
	0.9	0.158538	0.158455	0.158434
No interaction (target budget)	0.1	0.002133	0.002133	0.002133
	0.3	0.011611	0.011611	0.011611
	0.5	0.029588	0.029588	0.029588
	0.7	0.065808	0.065808	0.065808
	0.9	0.158250	0.158250	0.158250

C.4 FIXED-TARGET RECONSTRUCTION GEOMETRY ON A TWO-SPIRAL TOY

The synthetic distribution is a mixture of 64 isotropic Gaussians, with 32 centers on each of two spiral arms and component standard deviation 0.045. Both models are trained for 20,000 updates with seed 42 using four-layer, width-256 MLPs and batch size 512. For each of six fixed component centers (three evenly spaced radial positions per arm), we draw 1,024 independent noise pairs at $t = s = 0.4$. All methods receive the same noise realizations. Output averaging uses the mean of two predictions from the single-state model.

We subtract the fixed target and rotate each error into its radial and tangential coordinates. Figure 8 pools all six targets equally; Figure 9 separates the three radial regions. Dots show empirical mean errors, and ellipses show covariance contours with semi-axes twice the corresponding standard deviations. This conditional reconstruction probe distinguishes mean-error reduction from changes in error spread. Averaging has lower MSE for the inner targets; DualFM has lower MSE for the middle and outer targets and for the equally pooled comparison.

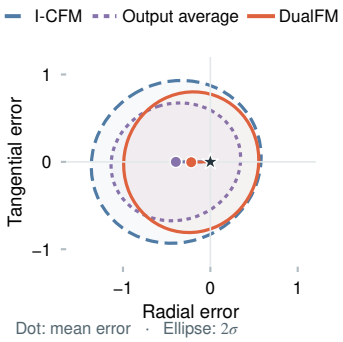

Figure 8: **Pooled fixed-target reconstruction errors.** Means and 2σ covariance ellipses over all six targets, in radial–tangential coordinates. The star marks zero error.

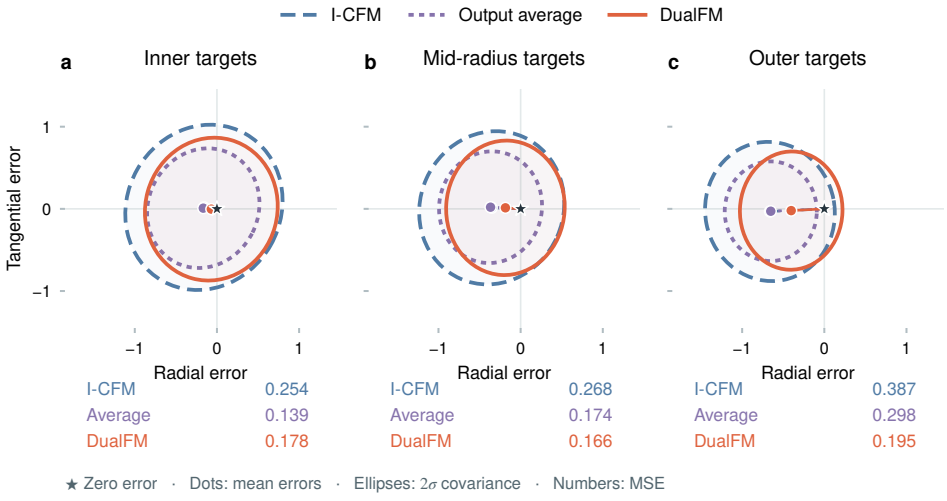

Figure 9: **Fixed-target error geometry by radius.** All three predefined target regions are retained. Numbers below each panel are reconstruction MSEs; the star marks zero error.

C.5 OUTPUT-BRANCH AGREEMENT DURING GENERATION

We evaluate the main ImageNet32 DualFM checkpoint (training seed 42, 600k updates, EMA weights) on 50,000 joint trajectories at each of 10, 20, 50, 100, and 1000 uniform Euler steps. Each trajectory yields X_0, H_0, and their mean; all three outputs therefore share the same initial noise pair and numerical integration. We apply the main CleanFID legacy-tensorflow pipeline and ImageNet32 training reference statistics separately to each output. This evaluation uses eight GPUs, batch size 256 per GPU, and FP32 without TF32. Sample-indexed noise draws with evaluation seed 2025 keep the initial pairs fixed across budgets.

Table 15 reports the three FIDs and $\mathbb{E}[\|X_0 - H_0\|_d^2]$ in the unmodified model output coordinates. Branch FIDs differ by at most 0.0453, with endpoint MSE below 9×10^{-6} throughout. Averaging does not improve X's FID at any measured budget. The two branches therefore agree closely during generation as well as on interpolation states.

Table 15: **Both output branches have similar generation quality.** ImageNet32, one 600k EMA checkpoint, 50,000 joint trajectories per budget. FID is computed separately for each output; endpoint MSE is measured before clipping, in model coordinates.

NFE	FID: X_0	FID: H_0	FID: $(X_0 + H_0)/2$	Endpoint MSE ($\times 10^{-6}$)
10	19.345	19.299	19.438	8.902
20	12.667	12.654	12.740	8.562
50	10.027	10.030	10.084	8.953
100	9.237	9.265	9.291	8.879
1000	8.545	8.564	8.564	5.824

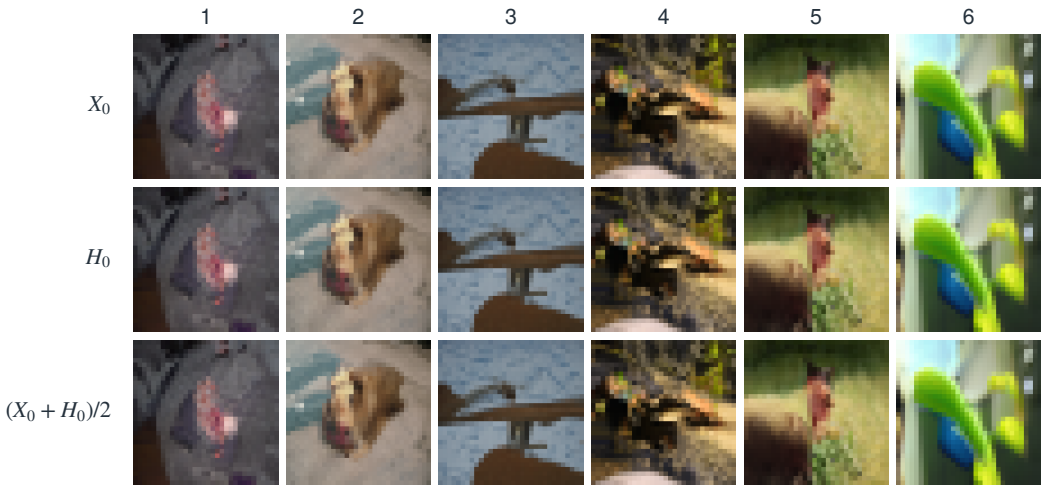

Figure 10: **Paired outputs from the same joint trajectories.** The first six samples at Euler-100, without selection. Each column shows X_0, H_0, and their mean for one initial noise pair. Images use the same fixed clipping and uint8 conversion as FID evaluation.

D ADDITIONAL GENERATED EXAMPLES

All examples use the main DualFM checkpoints (seed 42, 600k EMA), synchronized Euler-100, and FP32. Images are displayed without enhancement.

D.1 FIXED SAMPLE SOURCE, DIFFERENT AUXILIARY SOURCES

Figure 11 extends Figure 1 with one additional pair per dataset: the second generated pair (seed 3025). The two rows within each panel share the exact X_1 tensor and differ only in H_1. Noise is displayed on a fixed $[-3, 3]$ scale and outputs on $[-1, 1]$, with clipping for display only. Sampling is deterministic given the full pair; these images illustrate how the auxiliary source changes the generated image at fixed X_1.

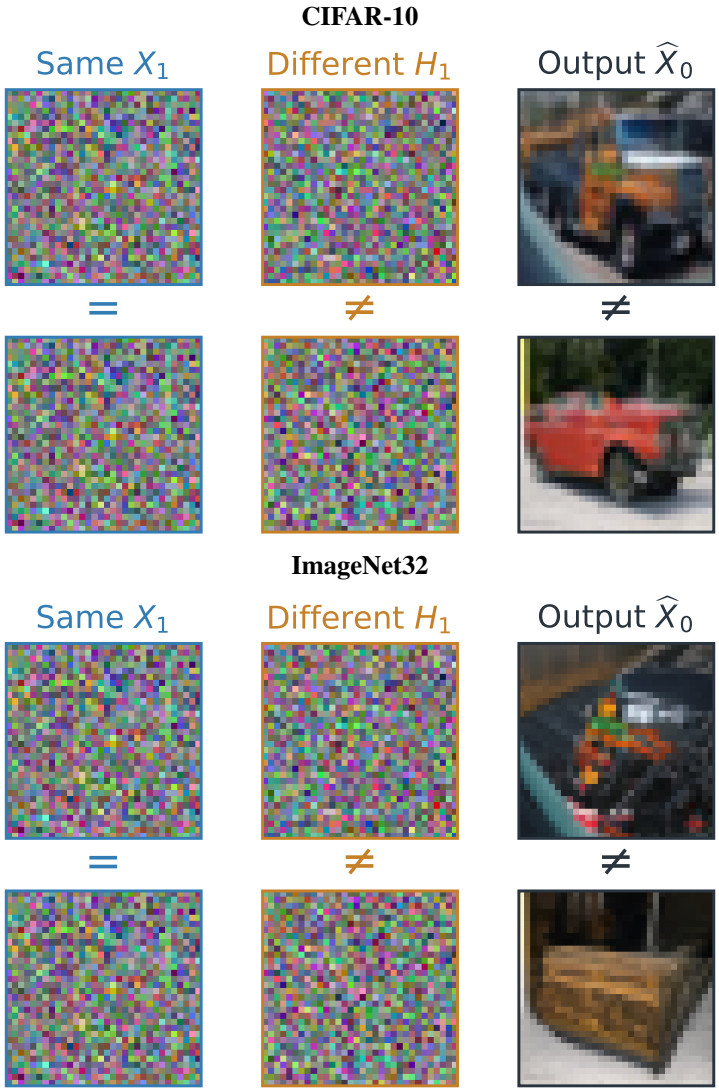

Figure 11: **Additional fixed-X_1 examples.** Columns show the sample source, auxiliary source, and returned image. The $=$ and $\neq$ markers identify identical X noise and different H noise across the two rows.

D.2 UNCURATED CIFAR-10 SAMPLES

Figure 12 shows 64 consecutive CIFAR-10 outputs in row-major order, using generation seed 2025 and the sampling protocol above. Native 32×32 outputs are enlarged with nearest-neighbor interpolation.

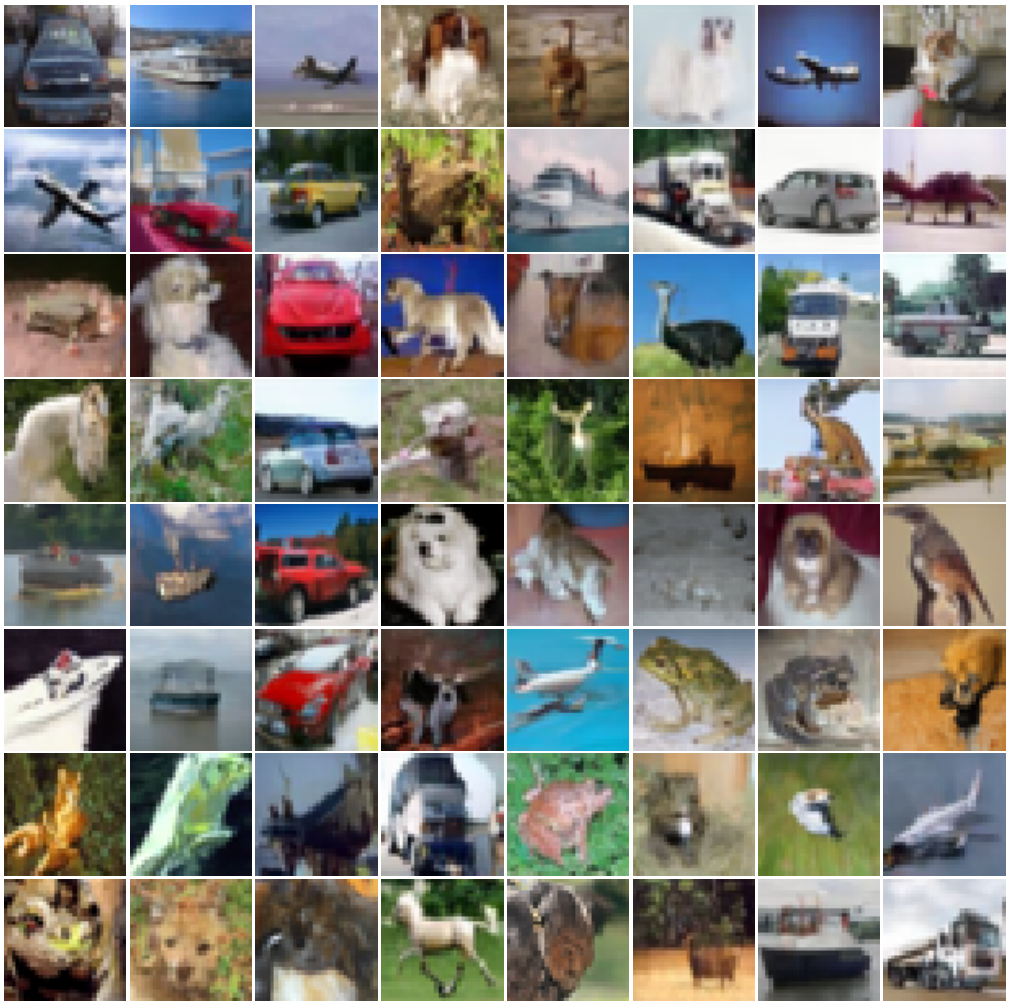

Figure 12: **Uncurated CIFAR-10 samples.** DualFM, 600k EMA, Euler-100.