TEMPORAL SOURCE COVARIANCE MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Which source covariance should a few-step flow-matching model use? We formulate source selection as a finite-step numerical design problem. For Gaussian population fields under independent coupling and linear interpolation, we prove that covariance matching uniquely minimizes endpoint Wasserstein error for every $N \geq 2$ uniform Euler steps, including non-commuting source covariances. The resulting covariance-only oracle has second-order excess risk near the optimum. Temporal Source Covariance Matching (TSCM) implements this principle with an offline, length-indexed Gaussian source estimated from within-channel temporal second moments and used in both training and sampling, without changing the architecture or objective. On HumanML3D, a controlled comparison of 16 sources with three training seeds each gives mean 24-step FIDs of 0.108 for TSCM, 0.135 for uncentered Toeplitz, and 0.497 for white noise, with source-specific constant guidance selected on validation. Oracle risk is associated with these guided test FIDs (log–log Pearson $r = 0.93$). Frozen-field controls also reveal learned-flow differences that persist under reference integration. Further analyses characterize solver-dependent calibration, conditional shared-source regret, and local covariance-deficit minimaxity beyond Gaussian targets. A separate guidance analysis identifies an early covariance bias; matched source replacements in other sequence generators assess empirical transfer.

1 INTRODUCTION

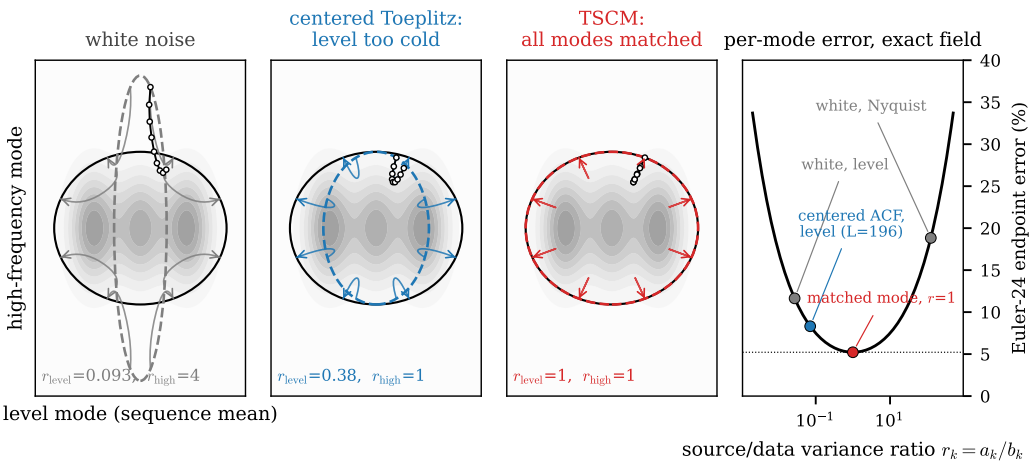

Figure 1: **Variance ratios determine mode-wise Euler error.** Left: decoupled level/high-frequency schematic, with data density in grey, a black covariance ellipse, dashed source contours, flow lines and 8-step Euler paths. Matching both directions reduces distortion; the data directions need not be decoupled. Right: exact 24-step relative endpoint error (Proposition 2), with projected HumanML3D variance ratios placed on the scalar curve for illustration.

Flow-matching models generate data by progressively transforming random noise into a structured sample. For multivariate sequences, this transformation must recover both within-step structure and

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.

dependence across time. Performing it with only a few sampling steps is important for efficient generation, but leaves limited room for numerical error. Most models start from independent Gaussian noise, which assigns equal variance to all normalized temporal patterns, from a sequence-wide offset to rapid frame-to-frame alternation. Real sequences can allocate very different amounts of variation to these patterns. This raises a practical question: can source covariance reduce the numerical burden of generation at a fixed sampling budget?

Prior work has explored temporally correlated or data/condition-adaptive sources (Kollovieh et al., 2025; Lee et al., 2022; Yang et al., 2026). Here we ask how to choose source covariance by minimizing finite-step numerical error. Intuitively, a temporal pattern given too much variance at initialization must be contracted, while one given too little must be expanded. In the flow we analyze, larger variance mismatches produce stronger dynamics, which a small number of sampling steps must resolve. Under independent source–target sampling and linear interpolation, we show for arbitrary positive-definite Gaussian covariances that matching uniquely minimizes the actual endpoint Wasserstein error after every $N \geq 2$ uniform Euler steps, even against non-commuting sources. Euler still leaves a nonzero matched-source error floor.

Temporal Source Covariance Matching (TSCM) instantiates this design principle through length-indexed within-channel temporal statistics. Its offline Gaussian source preserves sequence-level variation and fluctuations without a fixed kernel and is used unchanged in training and sampling, without an added module or changes to the architecture or objective.

On HumanML3D, a controlled comparison of 16 Gaussian sources uses the same DiT-M architecture, training protocol, and 24-step Euler sampler, with three training seeds per source and checkpoints and constant guidance selected on validation. Mean test FID is 0.108 for TSCM, 0.135 for uncentered Toeplitz, and 0.497 for white noise; the covariance-only oracle correlates with guided test FID across sources (log–log Pearson $r = 0.93$). Matched source replacements in other sequence generators test empirical transfer. Frozen-field and common-guidance controls examine how numerical residuals, learned flows, and guidance selection relate to the observed results.

The exact Gaussian uniform-Euler optimum is our central result; TSCM implements it on a block-diagonal temporal surrogate. Solver, conditional, and non-Gaussian analyses characterize related design objectives beyond this setting. Guidance induces a separate early covariance bias and motivates a complementary sampling intervention.

Closest prior work. TSFlow (Kollovieh et al., 2025) uses prescribed Gaussian-process kernels to introduce temporal dependence. PriorGrad (Lee et al., 2022) adapts diagonal Gaussian priors for speech synthesis, while CW-Gen (Yang et al., 2026) predicts condition-dependent means and per-timestep cross-variable covariances. TSCM instead estimates full within-channel temporal matrices offline, indexed by length rather than semantic condition. The distinction is also the optimization target: prior analyses address, among other benefits, training conditioning and data-to-prior divergence; we derive the finite-step Euler risk as a function of source covariance and solve the resulting optimization problem. For guidance, interval schedules (Kynkäänniemi et al., 2024) and stage-wise analyses (Jin et al., 2026) already motivate time-varying strength. Our analysis quantifies the initial translation and continuous-time covariance excess under independent coupling, connecting early guidance to sequence-level calibration. Appendix C gives the detailed comparisons.

2 Source covariance and numerical error

We ask which source covariance minimizes finite-step Euler error. Let the centred target x_1 have covariance $B \succ 0$, and let $x_0 \sim \mathcal{N}(0, A)$, $A \succ 0$, be independent of x_1. Under the linear interpolation $x_t = (1-t)x_0 + tx_1$, the population-optimal field is $v^*(x,t) = \mathbb{E}[x_1 - x_0 \mid x_t = x]$. We isolate the part of this field determined solely by A and B, then reduce its dynamics to generalized source–target variance ratios and derive their exact finite-step Euler error.

Let

$$\Sigma_t = (1-t)^2 A + t^2 B, \qquad G_t = [tB - (1-t)A]\Sigma_t^{-1}.$$

Then

$$v^*(x,t) = G_t\, x + r_t(x), \qquad \mathbb{E}[r_t(x_t)] = 0, \quad \mathbb{E}[r_t(x_t)x_t^\top] = 0. \tag{1}$$

For any centred target, $G_t x$ is the best linear projection of the population field under the interpolation law; for a Gaussian target the residual vanishes (Appendix A.1). We first solve the Gaussian problem for general $A, B \succ 0$, then apply it to a temporal model that retains within-channel covariance.

2.1 Variance ratios and mode-wise dynamics

Whiten by the target covariance and diagonalize the source:

$$B^{-1/2} A B^{-1/2} = V \operatorname{diag}(r_k) V^\top, \qquad z = V^\top B^{-1/2} x.$$

Then each generalized mode has target variance 1, source variance r_k, and evolves independently as $\dot{z}_k = g(t; r_k) z_k$ (Appendix A.2).

Proposition 1 (mode-wise gain)**.** *For source–target variance ratio r,*

$$g(t; r) = \frac{t - (1-t)\, r}{(1-t)^2 r + t^2}, \qquad \Lambda(r) = \sup_t |g(t; r)| = \frac{\sqrt{r} + 1/\sqrt{r}}{2} = \cosh\!\big(\tfrac{1}{2} \log r\big). \tag{2}$$

Hence $\Lambda(r)$ is uniquely minimized at $r = 1$ and grows symmetrically with $|\log r|$: mismatch strengthens the modal dynamics, with $r < 1$ giving an earlier transition to expansion and $r > 1$ prolonged contraction. The r_k are generalized eigenvalue ratios, not generally ratios along fixed level or Fourier directions; those projections are used only as diagnostics in Section 3.

2.2 Euler endpoint error and oracle risk

Because the generalized modes decouple, uniform Euler reduces to an independent scalar recursion in each mode, whose endpoint error is available in closed form.

Proposition 2 (Euler endpoint error)**.** *For a mode with source–target variance ratio r, N uniform Euler steps give*

$$R_N(r) = \sqrt{r} \prod_{n=0}^{N-1} \left(1 + \frac{g(n/N; r)}{N}\right), \qquad |1 - R_N(r)| = \frac{1}{2N}\left[1 + \frac{\pi}{2}\Lambda(r)\right] + O(N^{-2}), \tag{3}$$

where $R_N(r)$ is the generated standard deviation relative to the target standard deviation.

We use the exact product in Eq. (3); the expansion only makes explicit the link between peak gain and Euler error (Appendix A.3). Thus larger variance mismatch produces larger finite-step error. Even at the matched ratio $r = 1$, Euler retains a positive discretisation floor: the exact relative error is 5.2% at $N = 24$ (Figure 1).

Theorem 1 (Gaussian endpoint covariance optimum)**.** *Let $P_{N,A,B}$ be the N-step uniform-Euler endpoint of the exact Gaussian population field with independent source $\mathcal{N}(0, A)$ and target $\mathcal{N}(0, B)$, where $A, B \succ 0$. For every $N \geq 2$,*

$$W_2(P_{N,A,B}, \mathcal{N}(0, B)) \geq [1 - R_N(1)]\sqrt{\operatorname{tr} B}, \quad \textit{with equality iff } A = B. \tag{4}$$

The optimum includes non-commuting competitors.

The scalar Euler product satisfies $R_N(r) \leq R_N(1) < 1$, strictly unless $r = 1$ (Theorem A.1); hence $\operatorname{Cov}(P_{N,A,B}) \preceq R_N(1)^2 B$ and has strictly smaller trace when $A \neq B$. The trace lower bound for W_2 then gives the unique endpoint optimum (Corollary A.2).

Aggregating the mode-wise errors with target-variance weights gives a computable oracle risk (write $S_N(A)$ when B is fixed):

$$S_N(A; B)^2 = \sum_k w_k \big[1 - R_N(r_k)\big]^2, \qquad w_k = v_k^\top B v_k. \tag{5}$$

Proposition 3 (oracle risk)**.** *$S_N(A; B)$ equals the endpoint W_2 distance between the N-step Euler output of the Gaussian population field and $\mathcal{N}(0, B)$ when A commutes with B, and upper-bounds it otherwise.*

The same scalar optimum uniquely minimizes $S_N(A; B)$. If $e^{-\delta} B \preceq A \preceq e^{\delta} B$, its excess over the matched floor is $O(\delta^2)$ (Appendix A.5). For temporal experiments we take the explicit block-diagonal Gaussian model $B_L = \operatorname{diag}(B_{1,L}, \ldots, B_{C,L})$ and channel-independent sources, so Eq. (5) sums over channels. Restricting only the source does not imply this target choice (Appendix A.6). Section 2.5 tests the oracle risk against trained-model FID.

2.3 BEYOND UNIFORM EULER AND GAUSSIAN TARGETS

Covariance shape and solver-dependent scale. For Gaussian fields whose generalized modes share a continuous scalar solver response, an actual endpoint W_2 optimum can be chosen as $A = \kappa_* B$: a scalar calibration preserves the target covariance shape. Theorem A.3 applies on connected admissible ratio intervals where the scalar minimum is attained, for fixed paths and grids and a common aligned endpoint correlation. One-step Heun illustrates the change: under independent coupling its optimum is $\kappa_* = 4$. Further solver and grid analyses are in Appendix A.8.

Matching cancels the leading endpoint error. Let $\alpha(t) = \beta(1-t)$ define a smooth reflection-symmetric path, and use a reflection-symmetric deterministic grid of maximum step $h_{\max}$. For any fixed even-order p explicit Runge–Kutta method,

$$|R_{\mathcal{M},\mathcal{T}}(e^s) - 1| \le C\big(|s| h_{\max}^p + h_{\max}^{p+1}\big), \qquad |s| \le \delta_0. \tag{6}$$

At the match, time antisymmetry cancels the leading endpoint error, giving third-order Heun and fifth-order RK4 accuracy under these conditions (Theorem A.4). Exact-field convergence checks are in Appendix A.14.

A sharp local guarantee without Gaussian targets. For any target distribution with covariance B, a population Euler step started from the true interpolation law loses covariance $h^2 \mathbb{E}\,\mathrm{Cov}(x_1 - x_0 \mid x_t)$. With conditioning, the deficit can only decrease. Among all target–condition laws with the same marginal covariance, a Gaussian target with an uninformative condition attains the largest integrated, covariance-normalized deficit, equal to $d + \frac{\pi}{2}\sum_k \Lambda(r_k)$ (Theorem A.7). Matching uniquely minimizes this sharp minimax criterion for the normalized local covariance deficit integrated along the interpolation path.

2.4 SHARED-SOURCE DESIGN UNDER CONDITIONING

At fixed A, the marginal B defines the best condition-independent affine projection under the joint interpolation law (Proposition A.4). For one channel, assume $x_1 \mid Y = y \sim \mathcal{N}(\mu_y, B_y)$ and $x_0 \sim \mathcal{N}(0, A)$ independently of (x_1, Y), with $B_y \succeq 0$, $W = \mathbb{E}B_Y$, $V = \mathrm{Cov}(\mu_Y)$, $B = W + V \succ 0$, and $\mathbb{E}\|B_Y\|_F^2 < \infty$. Let $\mathcal{R}_N(A)$ be the average squared conditional endpoint W_2 of the exact unguided field with one shared source A and $N \ge 2$ uniform Euler steps. With $\mathcal{H} = \mathbb{E}[(B_Y - W)^2]$,

$$0 \le \mathcal{R}_N(B) - \inf_{A \succ 0} \mathcal{R}_N(A) \le \kappa_N \big[\,\mathrm{tr}(VB^{-1}V) + \mathrm{tr}(B^{-1}\mathcal{H})\big], \tag{7}$$

Here $\kappa_N = O(N^{-2})$. The bound compares B with the best shared source, allows non-commuting or singular B_Y, and separates mean variation from covariance heterogeneity (Theorem A.6). *Marginal-statistics design.* TSCM uses pooled marginal temporal statistics, directly estimable without fitting conditional moments. Equation (7) bounds the regret of this shared-source choice; it does not make B a universal conditional optimum. W is exactly optimal for homoscedastic Gaussian conditionals, while heterogeneous families can prefer another shared source. Appendix A.11 develops conditional-moment alternatives and their estimation requirements; Appendix A.11.7 reports caption-cluster budgets. Guidance bias is separate (Section 4).

2.5 EMPIRICAL EVALUATION OF THE ORACLE RISK

We test whether the covariance-only oracle predicts the behavior of trained models. We train 16 Gaussian sources on HumanML3D, spanning TSCM, white noise, the three TSFlow kernels, data-driven Toeplitz sources, and controlled level and temporal-structure variants, using the same DiT-M (Peebles & Xie, 2023; Ma et al., 2024), training recipe, and 24-step Euler sampler. Each source is trained with three independent seeds with the same checkpoint-search budget; a common discrete descent procedure selects each source's guidance scale on validation (Appendix B.2, Figure 7). We report mean $\pm$ SD across training seeds, with each seed's FID averaged over 20 test generations. Across three training seeds, guided FID is 0.497 ± 0.070 (white), 0.135 ± 0.002 (uncentered Toeplitz), and 0.108 ± 0.007 (TSCM; mean $\pm$ SD). Full source definitions and results are in Appendix B.3 (Table 8).

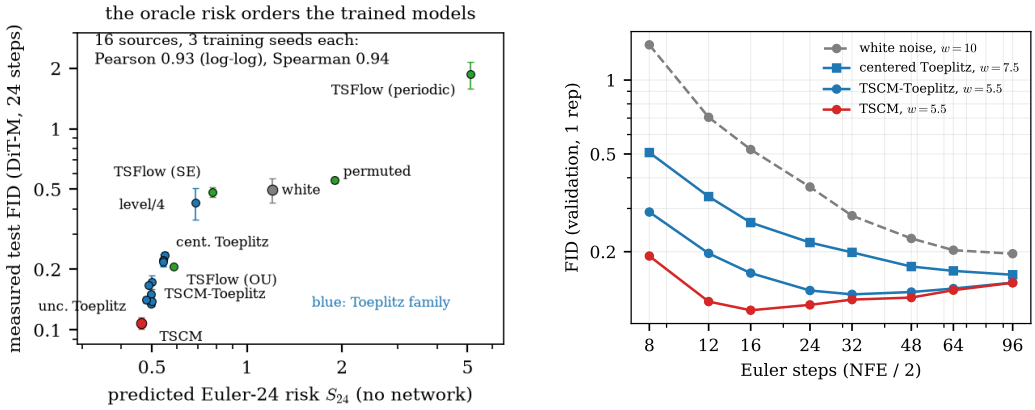

Figure 2: **Oracle risk and sampling steps.** **(a)** Oracle risk versus validation-selected guided test FID across 16 sources (log–log Pearson 0.93); checkpoints and constant guidance are selected on validation, and bars show SD across three training seeds. **(b)** Validation FID versus Euler steps on frozen checkpoints: white noise requires about 55 steps to match centered ACF at 24, while well-matched sources achieve better FID with fewer steps.

Does the oracle track trained-model quality? Across 16 sources, S_{24} has log–log Pearson correlation 0.93 with guided test FID under the common validation-selection protocol. This is an empirical association, not a theoretical bound on FID.

Does the association survive stricter controls? With final epoch-400 EMA checkpoints fixed for all 16 sources and three training seeds per source, S_{24} has log–log Pearson correlation 0.83 with the unguided Euler-24–RK4-192 endpoint residual (leave-one-source-out 0.80–0.88). Separately, a common-guidance scan over ten Toeplitz variants plus TSCM, using one fixed final-EMA checkpoint per source, gives log–log Pearson correlation 0.84 with Euler-24 FID at $w = 5.5$ (leave-one-source-out 0.78–0.86).

Where is the association less reliable? In that 11-source scan, the FID correlation is negative at $w = 1$ and $w = 10$. The within-family unguided residual association is sensitive to one low-level-power source. Thus the tested association at moderate guidance does not establish a guidance-invariant ranking of near-matched sources, nor separate guided continuous-flow and discretization effects (Appendix B.10).

Is source–data proximity enough? At $L = 64$, the periodic source is closer to validation windows than OU and white noise under empirical W_2 and energy distance, yet has worse 24-step FID (Appendix B.6, Table 10).

Can more sampling steps compensate for source mismatch? White noise needs about 55 steps to reach the centered-ACF model's 24-step FID, consistent with the oracle's predicted 2–2.6$\times$ increase in resolution, and at 96 steps still underperforms TSCM-Toeplitz at 24 (Figure 2b). Frozen-field diagnostics examine the remaining integration and learned-flow differences (Section 5, Appendix B.9).

3 TSCM: A LENGTH-INDEXED EMPIRICAL SOURCE

Theorem 1 identifies covariance matching as the unrestricted Gaussian uniform-Euler design target. TSCM applies this principle to a block-diagonal temporal surrogate, retaining dependence across time within each channel. For each supported length, we estimate temporal second moments from training windows and use their Cholesky factors in both training and sampling. The source is $q_L = \mathcal{N}(0, \bar{B}_L)$ with $\bar{B}_L = \mathrm{diag}(B_{1,L}, \ldots, B_{C,L})$.

Temporal-block approximation. The restriction to temporal blocks is a tractable estimation and storage choice, not a consequence of the unrestricted optimum. At $L = 196$ on HumanML3D, one dense float32 full covariance needs about 10.6 GB, versus 682 MB for TSCM's entire length-indexed factor table (Appendix B.5). Only $8{,}620$ eligible training windows are available at that length: the full empirical second moment has rank at most $8{,}620$ in $51{,}548$ dimensions, while each 196×196 temporal block uses all these windows (Appendices A.5 and B.7).

The blocks do not prescribe cross-channel dependence. Among zero-mean distributions with these prescribed within-channel temporal covariances, this block-diagonal Gaussian is the maximum-entropy choice (Appendix A.15.1). Appendix A.6 quantifies the surrogate's cost when the target has cross-channel covariance.

3.1 TSCM: FULL EMPIRICAL TEMPORAL COVARIANCE

For each supported length L in the training crop law and channel c, we draw one random L-frame window from each of the n_L training clips with at least L frames, in standardised data coordinates and without subtracting the window's own temporal mean. Window positions are drawn once with a fixed seed. Subtracting each window's temporal mean would remove its constant component and discard the between-window level variation that the source should retain. We estimate regularized second moments; their covariance interpretation uses the centered-data assumption:

$$\widehat{S}_{c,L} = \frac{1}{n_L} \sum_{i=1}^{n_L} x_c^{(i)} x_c^{(i)\top}, \qquad \widehat{A}_{c,L} = (1 - \rho_{c,L})\, \widehat{S}_{c,L} + \rho_{c,L}\, \mu_{c,L} I, \qquad x_{0,c} = \mathrm{Chol}(\widehat{A}_{c,L})\, \epsilon_c, \tag{8}$$

where $\epsilon_c \sim \mathcal{N}(0, I_L)$, $\mu_{c,L} = \operatorname{tr} \widehat{S}_{c,L}/L$, and $\rho_{c,L}$ is the Ledoit–Wolf shrinkage intensity (Ledoit & Wolf, 2004). In the model coordinates, residual mean energy is 0.24–1.77% of the second-moment trace on HumanML3D and at most 0.83% on TSFlow's eight datasets (Table 12, Appendix B.18). For D channels and supported lengths $\mathcal{L}$, we precompute the $D \times |\mathcal{L}|$ Cholesky factors and use the same source sampler during training and sampling, leaving the generator architecture and training objective unchanged.

3.2 TSCM-TOEPLITZ: A STRUCTURED APPROXIMATION

Let $u = \mathbf{1}/\sqrt{L}$ be the constant temporal direction and $H = I - uu^\top$ the projector onto fluctuations. The level power is $u^\top B_{c,L} u = L \operatorname{Var}(\text{window mean})$. Let C_c be the Toeplitz matrix of the sequence-centered training autocorrelation (ACF), and let $\hat{d}^{\mathrm{F}}_{c,L}$ be the level power of the Toeplitz matrix from the uncentered training ACF, computed by Eq. (A.85) in Appendix A.10. We define

$$A^{\text{TSCM-Toeplitz}}_{c,L} = HC_cH + \hat{d}^{\mathrm{F}}_{c,L} uu^\top. \tag{9}$$

This combines the centered-ACF fluctuation estimate with the uncentered-ACF level and sets the level–fluctuation cross block to zero. Appendix A.10 gives the full covariance decomposition and the Toeplitz level and positive-definiteness results.

3.3 WHAT STATIONARY SOURCES MISS

Figure 3 compares TSCM with structured and fixed-kernel stationary sources. The centered ACF captures temporal fluctuations but misses the level, the uncentered ACF only partly recovers long-window level power, and the TSFlow kernels miscalibrate either the spectrum or the level. Full TSCM instead tracks both across sequence lengths (Figure 3a,b; Appendix A.10). Figure 3(c) isolates length indexing in the covariance-only oracle; trained-model FIDs evaluate the complete TSCM source.

Independent-reference evaluation. To test whether TSCM's near-floor risk is an artifact of scoring against its own estimate, we use an independent held-out covariance $\widehat{B}_{\mathrm{ref}}$ and report the relative excess above the matched reference floor:

$$\frac{S_N(A; \widehat{B}_{\mathrm{ref}}) - S_N(\widehat{B}_{\mathrm{ref}}; \widehat{B}_{\mathrm{ref}})}{S_N(\widehat{B}_{\mathrm{ref}}; \widehat{B}_{\mathrm{ref}})}.$$

At $L = 64$, $N = 24$ on HumanML3D, TSCM is only 0.3% above the reference floor, versus 7.9% for TSCM-Toeplitz and 172% for white noise. Figure 3(c) shows the same pattern across lengths; Appendix B.7 gives the full source comparison and subsampling-budget results.

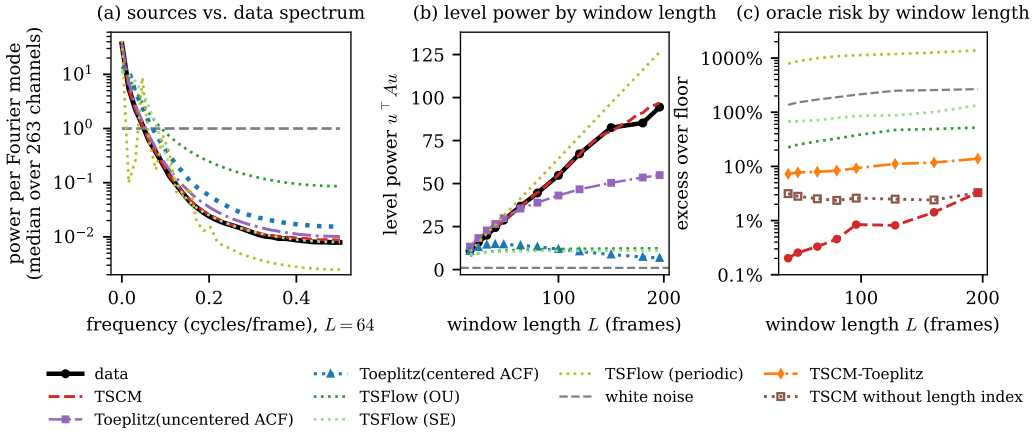

Figure 3: **The empirical source keeps what stationary estimates miss. (a)** TSCM matches the temporal spectrum; centered ACF captures the fluctuations but misses the level, while fixed-kernel sources miscalibrate the spectrum. **(b)** TSCM tracks level power across lengths, whereas stationary estimates increasingly miscalibrate long windows. **(c)** Against an independent covariance reference, TSCM remains near the oracle floor across lengths and has lower excess risk than Toeplitz approximations, TSFlow kernels, and white noise (Appendix B.7).

4 Level Calibration under Guidance

TSCM calibrates the source covariance, but strong classifier-free guidance (CFG) can immediately perturb this calibration. More Euler steps do not consistently improve FID after source matching (Figure 2b), suggesting a bias in the guided flow itself.

Let $v_w = v_u + w\Delta$, where $\Delta = v_c - v_u$, $\mu_y = \mathbb{E}[x_1 \mid y]$, $\mu = \mathbb{E}[x_1]$, and $V = \mathrm{Cov}_Y(\mu_Y)$. Under independent coupling, x_0 contains no condition information, giving $\Delta(x, 0, y) = \mu_y - \mu$. This translation changes covariance even under exact integration. Let $\Sigma_t = (1-t)^2 A + t^2 B$ denote the prescribed path covariance.

Proposition 4 (initial guidance covariance bias)**.** *Under the regularity assumptions of Appendix A.13, the population-optimal guided flow with constant w near $t = 0$ satisfies*

$$\mathrm{Cov}\left(X^{(w)}(t)\right) - \Sigma_t = t^2 w(w-1)V + o(t^2). \tag{10}$$

Projecting onto u replaces V by $u^\top V u$. For $w > 1$, CFG adds between-condition covariance that source calibration does not control, including excess level variance. This guidance-induced bias arises even under exact integration and is distinct from source-mismatch discretization error.

On the frozen TSCM DiT-M, the level component of Δ leaves the initial translation regime within a few steps (Figure 4a). RK4 trajectories follow the predicted $t^2 w(w-1)u^\top V u$ excess without a fitted constant for $t \le 0.02$ across $w = 2$–10 (Figure 4b; Appendix B.12). At the endpoint, constant guidance yields level variances 0.78 ($w = 5.5$) and 0.91 ($w = 7.5$), versus 0.54 for real motions, while reducing guidance early brings the variance closer to the data (Figure 4c).

On one frozen TSCM-Toeplitz model, sub-stepping the same early guided field does not improve FID, consistent with a guided-flow contribution (Appendices B.9–B.12).

We therefore reduce guidance only during this initial regime, using w_0 for $t < T_0$ and w_1 afterwards, with (w_0, T_0, w_1) selected on validation. Additional pairs and schedule ablations are in Appendix B.13 (Table 15).

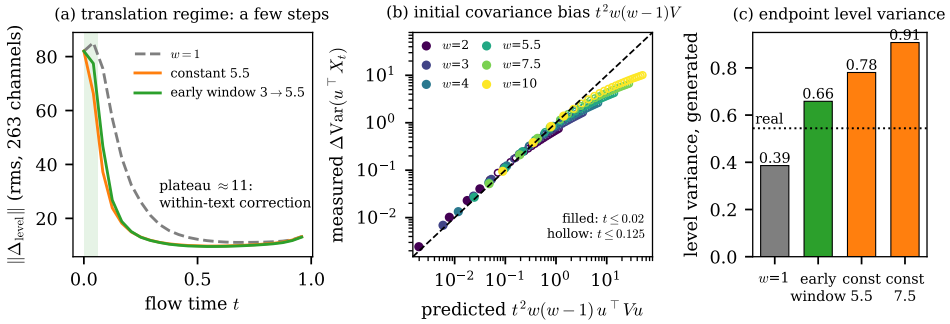

Figure 4: **Early guidance increases level spread.** **(a)** The level CFG difference leaves its initial regime within a few steps. **(b)** Early level-variance excess follows Proposition 4. **(c)** Reduced early guidance moves endpoint level variance toward the data.

5 Experiments

Matched source replacement. We use text-conditioned motion as a temporal and guidance testbed. The three-seed controlled study in Section 2.5 gives the main replicated evidence for source design. Table 3 extends the comparison to white/TSCM pairs with matched architecture, training budget, and validation selection within each generator. On HumanML3D (Guo et al., 2022), a separate single-run DiT-M comparison improves 24-step FID from 0.471 to 0.116 under selected constant guidance. MDM, Light-T2M, and DiT-L test transfer within other generator architectures; diffusion and temporally compressed latent results are detailed in Appendix B.

Complementary early guidance. Separately selected early-window schedules give DiT-M FIDs of 0.221 for white noise (Appendix B.13) and 0.076 for TSCM (Table 3). This is an additional sampling intervention, not a fixed-guidance factorial decomposition of the source effect.

Table 1 tests unguided unconditional time-series generation against training windows; Table 2 tests a single-caption video adaptation. Theory concerns Wasserstein and covariance objectives; FID/FVD measure learned distributions, while paired reference-flow residuals diagnose integration error (illustrations in Appendix B.19).

Table 1: **Unconditional time-series generation**: W_2 at 4 Euler steps against training windows. All results re-run the exact task, released code, and evaluation protocol of Table 1 in Kollovieh et al. (2025), changing only the source. Mean ± SD over three training seeds (Appendix B.18).

dataset	white noise	TSFlow (OU)	TSFlow (SE)	TSFlow (periodic)	TSCM-Toeplitz	TSCM
Electricity	6.64 ± 0.22	6.29 ± 0.30	6.52 ± 0.15	6.11 ± 0.19	6.31 ± 0.04	$\mathbf{6.00} \pm 0.11$
Solar	18.97 ± 1.72	18.18 ± 0.82	17.10 ± 0.76	10.97 ± 0.37	11.60 ± 0.50	$\mathbf{10.95} \pm 0.10$
Traffic	12.17 ± 0.18	12.04 ± 0.22	11.86 ± 0.14	11.52 ± 0.62	11.50 ± 0.08	$\mathbf{11.26} \pm 0.09$
UberTLC	40.46 ± 1.02	40.02 ± 0.26	40.24 ± 0.86	39.65 ± 0.20	39.92 ± 0.84	$\mathbf{39.31} \pm 0.55$
KDD-Cup	14.70 ± 0.09	13.68 ± 0.05	13.64 ± 0.04	15.04 ± 0.23	14.31 ± 0.10	$\mathbf{13.20} \pm 0.31$
M4-hourly	7.92 ± 0.18	7.75 ± 0.16	7.93 ± 0.21	7.07 ± 0.64	7.40 ± 0.15	$\mathbf{7.04} \pm 0.19$
Exchange	1.13 ± 0.18	1.02 ± 0.03	1.04 ± 0.05	1.32 ± 0.02	0.92 ± 0.03	$\mathbf{0.82} \pm 0.07$
Wiki2000	33.50 ± 0.10	33.05 ± 0.10	33.17 ± 0.02	33.03 ± 0.09	33.10 ± 0.09	$\mathbf{32.81} \pm 0.06$

Table 2: **Single-caption Sky Time-lapse adaptation.** FVD↓ on 512 test clips; Wan2.1 (Wan Team, 2025) fine-tunes change only the source. Δ: relative to white (Appendix B.17).

	50 steps		16 steps		8 steps	
method	FVD↓	Δ	FVD↓	Δ	FVD↓	Δ
+3 epochs, white prior	339.6	–	345.2	–	385.4	–
+3 epochs, TSCM prior	**228.7**	−33%	**215.2**	−38%	**201.9**	−48%

What remains after reference integration? At $w = 1$, final epoch-400 EMA models give RK4-192 FIDs of 2.36 ± 0.20 for white noise and 0.91 ± 0.13 for TSCM on 1024 test texts over three training seeds per source. Doubling reference resolution on 192 texts changes endpoints by less than 8×10^{-4} in relative RMS, so the learned unguided flows differ beyond Euler discretization error. Figure 5 shows their spectra and paired Euler residuals; Appendix B.10 defines and reports the frequency-resolved field-fitting diagnostic. These unguided final-checkpoint results use a different protocol from the validation-selected guided FIDs and do not additively decompose their gain.

Table 3: **Within-model HumanML3D/KIT-ML results.** All rows are our runs; white/TSCM pairs match architecture and budget, while EW changes guidance. $\pm$: 95% confidence interval over 20 test generations from a fixed checkpoint.

			R-precision↑			FID↓	Diversity↑
dataset	model	source / schedule	top-1	top-2	top-3		
HumanML3D	DiT-M	white	$0.489^{\pm.002}$	$0.686^{\pm.002}$	$0.790^{\pm.002}$	$0.471^{\pm.006}$	$9.877^{\pm.098}$
		TSCM	$0.530^{\pm.002}$	$0.725^{\pm.002}$	$0.823^{\pm.002}$	$0.116^{\pm.003}$	$9.742^{\pm.099}$
		TSCM + EW	$0.526^{\pm.002}$	$0.721^{\pm.002}$	$0.819^{\pm.002}$	$0.076^{\pm.002}$	$9.672^{\pm.098}$
	DiT-L	white	$0.558^{\pm.002}$	$0.753^{\pm.002}$	$0.844^{\pm.002}$	$0.345^{\pm.004}$	$9.991^{\pm.085}$
		TSCM	$0.568^{\pm.002}$	$0.759^{\pm.001}$	$0.847^{\pm.002}$	$0.047^{\pm.001}$	$9.656^{\pm.096}$
		TSCM + EW	$0.562^{\pm.002}$	$0.755^{\pm.002}$	$0.844^{\pm.002}$	$0.032^{\pm.001}$	$9.553^{\pm.102}$
KIT-ML	DiT-L	white	$0.488^{\pm.005}$	$0.728^{\pm.004}$	$0.844^{\pm.004}$	$0.172^{\pm.003}$	$11.288^{\pm.122}$
		TSCM	$0.475^{\pm.004}$	$0.700^{\pm.004}$	$0.829^{\pm.002}$	$0.120^{\pm.002}$	$11.184^{\pm.134}$
		TSCM + EW	$0.475^{\pm.004}$	$0.702^{\pm.004}$	$0.830^{\pm.002}$	$0.106^{\pm.002}$	$11.125^{\pm.135}$
HumanML3D	MDM (Tevet et al., 2023)	white	$0.331^{\pm.005}$	$0.489^{\pm.004}$	$0.622^{\pm.007}$	$0.537^{\pm.044}$	$9.359^{\pm.086}$
		TSCM	$0.523^{\pm.002}$	$0.720^{\pm.002}$	$0.817^{\pm.002}$	$0.380^{\pm.005}$	$10.242^{\pm.088}$
		TSCM + EW	$0.512^{\pm.002}$	$0.711^{\pm.002}$	$0.808^{\pm.001}$	$0.205^{\pm.004}$	$10.028^{\pm.088}$
	Light-T2M (Zeng et al., 2025)	white	$0.501^{\pm.003}$	$0.685^{\pm.002}$	$0.793^{\pm.002}$	$0.042^{\pm.002}$	$9.677^{\pm.107}$
		TSCM	$0.506^{\pm.002}$	$0.699^{\pm.001}$	$0.796^{\pm.001}$	$0.035^{\pm.001}$	$9.552^{\pm.093}$
		TSCM + EW	$0.503^{\pm.001}$	$0.699^{\pm.002}$	$0.796^{\pm.002}$	$0.033^{\pm.001}$	$9.504^{\pm.093}$

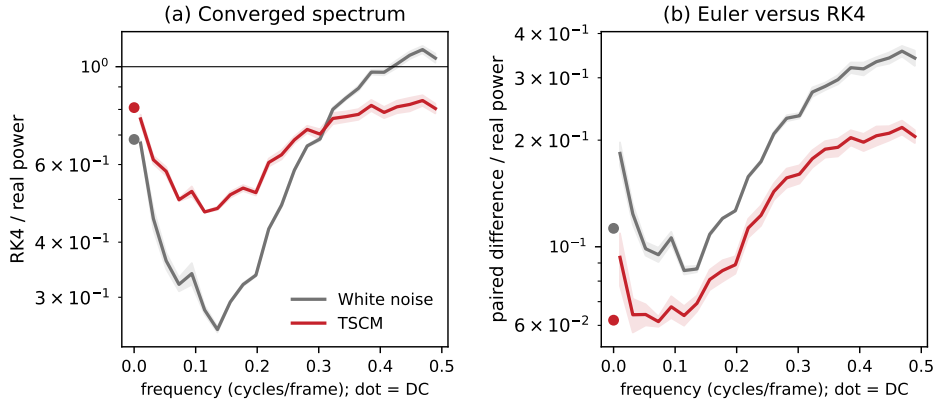

Figure 5: **Unguided DiT-M diagnostics.** RK4-192 spectrum (a) and paired Euler-24–RK4-192 residual (b), relative to real power; gray: white, red: TSCM. Three seeds, 1024 test texts each.

6 Conclusion and Future Work

We formulate source covariance selection as a finite-step numerical design problem and solve it exactly for Gaussian population fields under independent coupling, linear interpolation, and uniform Euler. TSCM applies this principle through an offline, length-indexed temporal approximation, supported by controlled sources and matched-generator experiments. Reference-flow and guidance diagnostics show that the Gaussian oracle does not fully determine learned-model quality and source calibration does not remove guidance-induced bias. The central implication is to design the source together with the finite-step generator. Future work can use conditional moments and tractable cross-channel statistics.

Reproducibility statement

Appendix B lists the model, training and evaluation settings, every source definition used in the controlled study, and the protocols of the oracle-risk, reference-flow, frequency, and guidance diagnostics. All proofs are in Appendix A. Code and the source statistics will be released.

AI use statement

The authors led the development of the central argument linking source–target variance ratios, the finite-step Euler product, and the optimality of covariance matching. Subsequent extensions, including conditional covariance budgets, spectral comparability, higher-order Runge–Kutta endpoint cancellation, and local guarantees beyond Gaussian targets, were developed through iterative interaction between the authors and generative AI tools, with both contributing to their formulation, derivation, and refinement. AI tools also assisted with literature searches, reference verification, manuscript revision, and LaTeX preparation. Numerical claims were cross-checked against experimental records and recomputed statistics, and bibliographic details were checked against original sources. The authors take responsibility for the final manuscript, including all mathematical claims, proofs, experimental results, and conclusions.

References

Shadab Ahamed, Eshed Gal, Md Shahriar Rahim Siddiqui, Simon Ghyselincks, Moshe Eliasof, and Eldad Haber. Preconditioned flow matching. *arXiv preprint arXiv:2603.02337*, 2026. URL https://arxiv.org/abs/2603.02337v2. Version 2.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ernst Hairer and Marino Zennaro. On error growth functions of Runge–Kutta methods. *Applied Numerical Mathematics*, 22(1–3):205–216, 1996. doi: 10.1016/S0168-9274(96)00032-3.

Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.22007.

Yuma Koizumi, Heiga Zen, Kohei Yatabe, Nanxin Chen, and Michiel Bacchiani. SpecGrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping. In *Interspeech*, 2022. URL https://www.isca-archive.org/interspeech_2022/koizumi22_interspeech.html.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with gaussian process priors for probabilistic time series forecasting. In *International Conference on Learning Representations*, 2025. arXiv:2410.03024.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2403.13117.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *International Conference on Learning Representations*, 2022.

Heng Li, Xiaotong Lin, Ling-An Zeng, Yulei Kang, Shuai Li, and Jian-Fang Hu. MotionHiFlow: Text-to-motion via hierarchical flow matching. In *CVPR*, 2026. arXiv:2604.23264.

Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. FineDance: A fine-grained choreography dataset for 3D full body dance generation. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

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.

Jinseong Park, Seungyun Lee, Woojin Jeong, Yujin Choi, and Jaewook Lee. TimeBridge: Better diffusion prior design with bridge models for time series generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1135–1146, 2026. doi: 10.1145/3770854.3780195.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

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.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *CVPR*, 2018.

Kaixing Yang, Xulong Tang, Ziqiao Peng, Xiangyue Zhang, Puwei Wang, Jun He, and Hongyan Liu. FlowerDance: MeanFlow for efficient and refined 3D dance generation. *arXiv preprint arXiv:2511.21029*, 2025. to appear at ECCV 2026.

Yanfeng Yang, Siwei Chen, Pingping Hu, Zhaotong Shen, Yingjie Zhang, Zhuoran Sun, Shuai Li, Ziqi Chen, and Kenji Fukumizu. Conditionally whitened generative models for probabilistic time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GG01lCopSK.

Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. Light-T2M: A lightweight and fast model for text-to-motion generation. In *AAAI Conference on Artificial Intelligence*, 2025.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, pp. 14730–14740, 2023.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Yifei Zhu, Mingyi Shi, Yangyang Cai, Miao Cheng, Yoshifumi Kitamura, and Taku Komura. MoRAE: Flow-friendly self-supervised latents for text-to-motion generation. *arXiv preprint arXiv:2607.29180*, 2026.

CONTENTS OF THE APPENDIX

Overview: Source covariance as numerical design **15**

A Proofs and theoretical extensions **16**

A.1 Exact decomposition of the optimal field 16

A.2 Mode-wise gain (Proposition 1) 17

A.3 Euler endpoint error (Proposition 2) 17

A.4 Oracle risk (Proposition 3) and the exact uniform-Euler optimum 18

A.4.1 The Wasserstein identity and bound 18

A.4.2 The Euler product as a product of cosines 18

A.4.3 Strict log-concavity in the log-ratio 18

A.4.4 The positive tridiagonal form 19

A.5 Stability under approximate matching and the estimator error of TSCM 20

A.6 Cost of the block-diagonal surrogate 21

A.7 Explicit Runge–Kutta bounds; grids, paths and couplings 22

A.7.1 A source-dependent resolution scale 22

A.7.2 A one-step and endpoint theorem for explicit Runge–Kutta methods 22

A.7.3 Optimal grids for a single Euler mode 23

A.7.4 Smooth interpolation schedules 24

A.7.5 Coupled endpoints and the optimal-transport boundary 24

A.8 Covariance shape, scalar calibration, and endpoint cancellation 25

A.8.1 A Gaussian family with aligned endpoint correlations 25

A.8.2 An actual Wasserstein optimum with the target covariance shape 25

A.8.3 Endpoint cancellation for even-order Runge–Kutta methods 26

A.8.4 Heun: an explicit leading term and a calibrated scale 27

A.8.5 Joint optimization of covariance and a shared Euler grid 28

A.9 Proportional covariances and the transport distance 28

A.10 Positive definiteness and the level of a Toeplitz autocorrelation 29

A.11 Conditional flows and the marginal covariance 29

A.11.1 Which covariance describes which field 30

A.11.2 A counterexample to universal marginal matching 30

A.11.3 Average conditional error with a single shared source 30

A.11.4 The shared covariance that minimizes the upper bound 32

A.11.5 Why a conditional heterogeneity budget is necessary 33

A.11.6 A separate worst-mode transfer under spectral comparability 33

A.11.7 The conditional budget on HumanML3D 34

A.11.8 Held-out caption-cluster second-moment transfer 34

A.12 A sharp local covariance guarantee beyond Gaussian targets 36

A.12.1 An exact one-step identity and the effect of conditioning 36
A.12.2 Minimax calibration with only marginal second moments 37
A.12.3 A correlated extension of the local minimax budget 38
A.13 Guidance (Proposition 4) 38
A.13.1 The initial translation 38
A.13.2 The continuous-time early-guidance covariance expansion 38
A.13.3 The first Euler step as a discrete counterpart 39
A.13.4 How fast the state becomes informative 39
A.13.5 A one-dimensional Gaussian diagnostic 39
A.14 Reproducible checks on exact population fields 40
A.15 Numerical robustness under covariance shift 40
A.15.1 Source support and second-order covariance-shift cost 40
A.15.2 The minimax fixed source under unknown covariance 41
A.15.3 Test-condition averages and learned-field error 42

B Experimental details **43**
B.1 Training recipe 43
B.2 Evaluation protocol 44
B.3 The sources 44
B.4 Training-seed variation of the controlled study 45
B.5 Cost of the source 45
B.6 The oracle risk and its correlation with FID 47
B.7 Estimation stability of the TSCM statistics 48
B.8 The cross block 50
B.9 Reference-flow experiment on frozen models 50
B.10 Frozen-field source and frequency diagnostics 52
B.11 Higher-order solvers at equal cost 55
B.12 Guidance trajectories 56
B.13 Early-window guidance: pairs and variants 56
B.14 FineDance 57
B.15 Diffusion models: x_0-prediction DiT-M and MDM 57
B.16 Latent model 59
B.17 Sky Time-lapse with Wan2.1 59
B.18 TSFlow unconditional generation: details and 16-step results 60
B.19 Illustrative samples 60

C Related work **64**

OVERVIEW: SOURCE COVARIANCE AS NUMERICAL DESIGN

The central contribution of this work is to make source covariance an explicit numerical design variable for finite-step generation. Even when the corresponding population vector field is available exactly, the source can impose a numerical burden that a limited sampling budget must resolve. We formulate this dependence as an optimization problem, derive a training-free criterion for evaluating source covariances, and realize its solution through Temporal Source Covariance Matching (TSCM). The appendix develops the mathematical structure behind this perspective and the evidence connecting it to learned generators.

An exact finite-step design principle. Appendices A.1–A.4 show how generalized source–target variance ratios determine the covariance-driven dynamics and their exact Euler endpoint error. The analysis exposes a geometric cosine-product representation and strict log-concavity in the log variance ratio. For general positive-definite Gaussian covariances under independent coupling and linear interpolation, matching uniquely minimizes actual endpoint Wasserstein error for every $N \geq 2$ uniform Euler steps, including non-commuting competitors. Applying this result to the block-diagonal temporal model gives TSCM its design target. This gives covariance matching an exact finite-budget justification, rather than only a qualitative argument for smoother dynamics. The analysis also separates the matched-source discretization floor from the additional cost of covariance mismatch.

A computable criterion and a practical realization. The resulting oracle risk evaluates a candidate source directly from source and data covariances, without training a generator. TSCM implements the identified target through offline, length-indexed temporal statistics, used consistently during training and sampling without changing the generator architecture or loss formulation. Appendix A.5 establishes second-order excess risk near the optimum, connecting approximate covariance matching to numerical accuracy; Appendices A.10 and B.7 examine structured approximations and finite-data estimation. Crucially, the criterion has empirical predictive value: in the controlled HumanML3D study, its log–log Pearson correlation with test FID is 0.93 across 16 sources, each trained with three seeds. Together with independent-reference covariance checks and source-replacement experiments in motion, dance, video, and time series, this links the analytical design principle to practical generative modeling.

Structure beyond a single solver. Appendices A.7–A.9 reveal how source design interacts with the numerical scheme. In the Gaussian modal families analyzed here, an endpoint-optimal source can retain the target covariance shape while using a solver-dependent scalar calibration. Matching also cancels the leading endpoint error for even-order explicit Runge–Kutta methods on reflection-symmetric paths and grids, yielding third-order Heun and fifth-order RK4 endpoint accuracy. Joint source–grid optimization explains the role of covariance shape: equal generalized variance ratios allow all modes to share an optimal angular grid, while a common scale can be absorbed by step placement. These results identify distinct roles for source shape, source scale, and discretization, extending the numerical perspective beyond the uniform-Euler optimum.

Quantitative principles for conditional and non-Gaussian generation. Appendix A.11 develops a shared-source analysis for Gaussian conditional targets. It bounds the cost of using the marginal covariance relative to the best shared source through two interpretable quantities: conditional-mean variation and covariance heterogeneity. Appendix A.12 establishes a complementary, distribution-free result for an integrated, covariance-normalized local Euler deficit. Among target distributions with the same marginal covariance, a Gaussian target attains the worst case, and covariance matching uniquely minimizes this worst-case value. The resulting conditional regret bound and sharp local minimax criterion provide two further mathematical foundations for covariance-based source design.

Separating source calibration from guidance bias. Appendix A.13 identifies another mechanism relevant to few-step generation: under independent coupling, strong classifier-free guidance produces an early covariance excess even with exact integration. Its leading term is $t^2 w(w-1)\,\mathrm{Cov}_Y(\mu_Y)$, where $\mu_Y = \mathbb{E}[x_1 \mid Y]$ and $w > 1$ is the guidance strength. This separates guidance-induced distributional bias from source-mismatch discretization error and motivates the early-window schedule evaluated through frozen-model diagnostics. Source design and guidance calibration therefore address distinct mechanisms, rather than two interchangeable ways of increasing numerical resolution.

Appendix A.14 provides direct checks on exact population fields. Appendix A.15 extends source optimality to covariance uncertainty sets and separates numerical robustness from semantic extrapolation. Appendix B documents the experimental protocols and additional results, and Appendix C situates the contributions within related work. Taken together, the analysis and experiments support a shift from specifying a source independently of sampling to designing its covariance for the finite-step computation that generates samples. TSCM is a concrete realization of this perspective for temporal structure.

A Proofs and theoretical extensions

This section provides the proofs and theoretical extensions supporting Sections 2–4. Unless stated otherwise, the results concern exact population fields under the setting below.

Setting. Unless a subsection explicitly specifies a correlated coupling or singular conditional covariance, $x_0 \sim \mathcal{N}(0, A)$ with $A \succ 0$ is independent of (x_1, y); x_1 is centred with finite covariance $B \succ 0$; $x_t = (1-t)x_0 + tx_1$; and expectations conditional on x_t are taken under this interpolation. The Gaussian endpoint optimum permits arbitrary positive-definite A, B on the full state space. One-channel statements used for TSCM are aggregated over its explicit block-diagonal Gaussian surrogate $\bar{B} = \mathrm{diag}(B_1, \dots, B_C)$ with block-diagonal sources. We write

$$D(t) = (1-t)^2 r + t^2, \qquad g(t;r) = \frac{t-(1-t)r}{D(t)}, \qquad \Lambda(r) = \frac{\sqrt{r} + 1/\sqrt{r}}{2},$$
$$R_N(r) = \sqrt{r} \prod_{n=0}^{N-1} \Big(1 + \frac{g(n/N; r)}{N}\Big). \tag{A.1}$$

A.1 Exact decomposition of the optimal field

Proposition A.1 (decomposition and score residual). *Let $\Sigma_t = \mathrm{Cov}(x_t) = (1-t)^2 A + t^2 B$ and $G_t = [tB - (1-t)A]\Sigma_t^{-1}$. Then the optimal field $v^*(x,t) = \mathbb{E}[x_1 - x_0 \mid x_t = x]$ satisfies*

$$v^*(x,t) = G_t\, x + r_t(x), \qquad \mathbb{E}[r_t(x_t)] = 0, \qquad \mathbb{E}[r_t(x_t) x_t^\top] = 0, \tag{A.2}$$

and $r_t \equiv 0$ when x_1 is Gaussian. For $0 < t < 1$, the Gaussian-source residual is a score correction relative to the Gaussian φ_{Σ_t} of the same covariance,

$$r_t(x) = \frac{1-t}{t}\, A\, \nabla \log \frac{p_t(x)}{\varphi_{\Sigma_t}(x)}. \tag{A.3}$$

Proof. With independent centred endpoints,

$$\mathrm{Cov}(x_t) = (1-t)^2 A + t^2 B = \Sigma_t, \qquad \mathbb{E}\big[(x_1 - x_0) x_t^\top\big] = tB - (1-t)A, \tag{A.4}$$

so the best linear predictor of $x_1 - x_0$ from x_t is $G_t x$. By the tower property $\mathbb{E}[v^*(x_t,t)x_t^\top] = \mathbb{E}[(x_1 - x_0)x_t^\top]$, hence $r_t(x) = v^*(x,t) - G_t x$ has $\mathbb{E}[r_t(x_t)] = 0$ and $\mathbb{E}[r_t(x_t)x_t^\top] = 0$. If x_1 is Gaussian the conditional expectation is linear and $r_t \equiv 0$. For a Gaussian source, differentiating the convolution density gives

$$\nabla \log p_t(x) = -\frac{1}{1-t} A^{-1}\, \mathbb{E}[x_0 \mid x_t = x], \qquad \text{hence} \qquad v^*(x,t) = \frac{x}{t} + \frac{1-t}{t} A \nabla \log p_t(x); \tag{A.5}$$

replacing p_t by φ_{Σ_t} in the second identity reproduces $G_t x$, which gives Eq. (A.3). □

A.2 MODE-WISE GAIN (PROPOSITION 1)

Proof of Proposition 1. Let $B^{-1/2}AB^{-1/2} = V\,\mathrm{diag}(r_k)V^\top$ and $T = B^{-1/2}V$, so that $T^\top BT = I$ and $T^\top AT = \mathrm{diag}(r_k)$. In $z = T^\top x$ the endpoints have covariances $\mathrm{diag}(r_k)$ and I; therefore $T^\top \Sigma_t T$ and $T^\top[tB-(1-t)A]T$ are both diagonal and

$$\widetilde{G}_t = T^\top G_t T^{-\top} = \mathrm{diag}\Big(\frac{t-(1-t)r_k}{(1-t)^2 r_k + t^2}\Big), \tag{A.6}$$

which is Eq. (2). The endpoints give $g(0;r) = -1$ and $g(1;r) = 1$ for every r, and g vanishes exactly once, at $t = r/(1+r)$. For the peak, $\partial_t g = 0$ is equivalent to

$$(1+r)D(t) = 2\big(t-(1-t)r\big)^2, \tag{A.7}$$

whose admissible roots are

$$t^* = \frac{r-\sqrt{r}}{1+r}\ \ (r>1), \qquad t^* = \frac{r+\sqrt{r}}{1+r}\ \ (r<1), \qquad |g(t^*;r)| = \frac{\sqrt{r}+1/\sqrt{r}}{2} = \Lambda(r) \tag{A.8}$$

in both cases, with $g(t^*) = -\Lambda$ (a contraction) for $r > 1$ and $g(t^*) = +\Lambda$ (an expansion) for $r < 1$. $\Lambda(r) = \cosh(\frac{1}{2}\log r)$ is symmetric in $\log r$ and equals 1 only at $r = 1$. □

Remark A.1. On HumanML3D under white noise, the projected variance ratios at the two ends of the spectrum are $1/37$ for the level direction and 125 for the Nyquist direction; read as mode ratios they give $\Lambda = 3.1$ and 5.6, against $\Lambda = 1$ for a mode at $r = 1$. Projected ratios coincide with generalised eigenvalues only for joint eigenvectors (Section 2.1).

A.3 EULER ENDPOINT ERROR (PROPOSITION 2)

Proof of Proposition 2. On the diagonal field each mode obeys $\dot z = g(t;r)z$, so uniform Euler gives $z_{n+1} = z_n\big(1 + g(n/N;r)/N\big)$ and $z_N = z_0\prod_{n<N}\big(1+g(n/N;r)/N\big)$. Since z_0 has variance r and the data mode has variance 1, the generated-to-data standard-deviation ratio is $R_N(r)$ of Eq. (3), and the W_2 distance between two centred Gaussians on a line is the difference of their standard deviations. For the first-order term note $g = \frac{1}{2}\partial_t \log D$, so the exact flow has $\exp\int_0^1 g\,dt = \sqrt{D(1)/D(0)} = 1/\sqrt{r}$, i.e. $R_\infty = 1$. Expanding the product with $g_n = g(n/N;r)$,

$$\log R_N = \tfrac{1}{2}\log r + \sum_{n<N}\Big[\frac{g_n}{N} - \frac{g_n^2}{2N^2} + O(N^{-3})\Big], \tag{A.9}$$

$$\sum_{n<N}\frac{g_n}{N} = \int_0^1 g\,dt - \frac{g(1)-g(0)}{2N} + O(N^{-2}) = -\tfrac{1}{2}\log r - \frac{1}{N} + O(N^{-2}), \tag{A.10}$$

by the Euler–Maclaurin correction of the left Riemann sum and $g(0) = -1$, $g(1) = 1$. Hence

$$\log R_N = -\frac{1}{N}\Big[1 + \frac{1}{2}\int_0^1 g^2\,dt\Big] + O(N^{-2}). \tag{A.11}$$

The identity $(t-(1-t)r)^2 = (1+r)D - r$ gives $g^2 = (1+r)/D - r/D^2$, and with $s = t/(1-t)$,

$$\int_0^1 \frac{dt}{D} = \int_0^\infty \frac{ds}{r+s^2} = \frac{\pi}{2\sqrt{r}}, \qquad \int_0^1 \frac{r\,dt}{D^2} = 1 + \frac{\pi}{2}\Lambda(r), \tag{A.12}$$

so $\int_0^1 g^2 dt = \frac{\pi}{2}\Lambda(r) - 1$ and $1 - R_N = \frac{1}{2N}\big[1 + \frac{\pi}{2}\Lambda(r)\big] + O(N^{-2})$. □

Remark A.2 (range of the expansion). The expansion holds for fixed r as $N \to \infty$ and is not uniform in r. At $N = 24$ the exact product and the first-order formula compare as follows (the exact values are those quoted in the main text):

r	1	5/3	100 or 0.01	$10^{\pm 4}$
exact $1 - R_{24}(r)$	5.22%	5.32%	17.2%	77.8%
first order $\frac{1}{48}[1 + \frac{\pi}{2}\Lambda(r)]$	5.36%	5.46%	18.6%	166%

The exact product is what the oracle of Eq. (5) evaluates.

A.4 Oracle risk (Proposition 3) and the exact uniform-Euler optimum

A.4.1 The Wasserstein identity and bound

Proof of Proposition 3. Uniform Euler on the Gaussian population field is an affine map, $x_N = \Pi x_0$ with $\Pi = \prod_n (I + G_{t_n}/N)$, so the output is Gaussian with covariance $\Pi A \Pi^\top$ and its endpoint W_2 distance to $\mathcal{N}(0, B)$ is the Bures distance. When A and B commute, Π, A and B are simultaneously diagonal and the squared Bures distance is $\sum_k b_k(1 - R_N(r_k))^2$, i.e. $S_N(A;B)^2$ with $w_k = v_k^\top B v_k = b_k$. In general $S_N(A;B)^2$ is the squared distance of the mode-wise pairing of the two Gaussians, and a pairing distance upper-bounds the optimal-transport distance. (On our Toeplitz matrices the two agree to four digits at $L = 64$ and within 1% at $L = 196$.) □

The remainder proves that, for every $N \geq 2$, the oracle risk is uniquely minimized at $A = B$ for general positive-definite covariances.

A.4.2 The Euler product as a product of cosines

Lemma A.1 (the Euler product)**.** *Let $P(t) = \big((1-t)\sqrt{r},\, t\big) \in \mathbb{R}^2$ and let θ_n be the angle subtended at the origin by the segment from $P(n/N)$ to $P((n+1)/N)$. Then*

$$R_N(r) = \prod_{n=0}^{N-1} \cos\theta_n, \qquad \sum_{n=0}^{N-1} \theta_n = \frac{\pi}{2}. \tag{A.13}$$

Consequently: (i) $R_1 \equiv 0$, and for $N \geq 2$, $0 < R_N(r) < 1$ for every $r > 0$; (ii) $R_N(r) = R_N(1/r)$; (iii) $R_N(r) \to 0$ as $r \to 0$ or $r \to \infty$; (iv) $R_2(r) = \sqrt{r}/(1+r)$, maximised uniquely at $r = 1$.

Proof. Write $c_n = (1 - t_n)\sqrt{r} + i\, t_n$ with $t_n = n/N$, so $D(t_n) = |c_n|^2$. Expanding,

$$D(t_n)\Big(1 + \frac{g(t_n; r)}{N}\Big) = (1-t_n)(1-t_{n+1})r + t_n t_{n+1} = \mathrm{Re}(c_n \bar{c}_{n+1}) = |c_n||c_{n+1}| \cos\theta_n, \tag{A.14}$$

hence $1 + g(t_n)/N = (|c_{n+1}|/|c_n|)\cos\theta_n$. The product telescopes to $(|c_N|/|c_0|)\prod_n \cos\theta_n = r^{-1/2}\prod_n \cos\theta_n$, which gives $R_N = \prod_n \cos\theta_n$. The polar angle of $P(t)$ increases from 0 to $\pi/2$, which gives the sum. (i) For $N = 1$, $\theta_0 = \pi/2$; for $N \geq 2$ every $\theta_n \in (0, \pi/2)$. (ii) For $1/r$, $P(t)$ is $r^{-1/2}$ times $P(1-t)$ for r with the two axes exchanged, which preserves angles and reverses their order. (iii) As $r \to \infty$, $\theta_{N-1} \to \pi/2$; use (ii) for $r \to 0$. (iv) $R_2 = \cos\theta_0 \sin\theta_0 = \frac{1}{2}\sin 2\theta_0$ with $\tan\theta_0 = 1/\sqrt{r}$. □

A.4.3 Strict log-concavity in the log-ratio

Theorem A.1 (strict log-concavity and the unique optimum)**.** *For every integer $N \geq 2$, $s \mapsto \log R_N(e^s)$ is strictly concave on $\mathbb{R}$ and even. Consequently*

$$0 < R_N(r) \leq R_N(1) < 1, \qquad R_N(r) = R_N(1) \iff r = 1, \tag{A.15}$$

R_N is strictly increasing on $(0,1)$ and strictly decreasing on $(1,\infty)$, and for every $N \geq 2$ the oracle satisfies

$$S_N(A;B) \geq \big[1 - R_N(1)\big]\sqrt{\mathrm{tr}\, B}, \qquad \textit{with equality iff } A = B. \tag{A.16}$$

For $N = 1$, $R_1 \equiv 0$ and every ratio gives the same error.

Proof. Evenness is Lemma A.1(ii). For $n = 0, \ldots, N$ let

$$D_n = (N-n)^2 r + n^2, \qquad Q_n = (N-n)(N-n-1)r + n(n+1), \tag{A.17}$$

so that the Euler multiplier of step n is $1 + g(n/N; r)/N = Q_n/D_n$. With

$$a_n = \frac{n^2}{(N-n)^2}\ (1 \leq n \leq N-1), \quad b_n = \frac{n(n+1)}{(N-n)(N-n-1)} = \sqrt{a_n a_{n+1}}\ (1 \leq n \leq N-2), \tag{A.18}$$

the product telescopes, up to an additive constant independent of s, to

$$\log R_N(e^s) = \frac{s}{2} + \sum_{n=1}^{N-2} \log(e^s + b_n) - \sum_{n=1}^{N-1} \log(e^s + a_n). \tag{A.19}$$

Define

$$f_n = \frac{\sqrt{r}\, n(N-n)}{D_n} > 0 \ \ (1 \le n \le N-1), \qquad c_n = \frac{D_n D_{n+1}}{Q_n^2} > 0 \ \ (1 \le n \le N-2). \tag{A.20}$$

Differentiating Eq. (A.19) twice gives the exact identity

$$\frac{d^2}{ds^2} \log R_N(e^s) = -\sum_{n=1}^{N-1} f_n^2 + \sum_{n=1}^{N-2} c_n f_n f_{n+1} = -f^\top K f, \tag{A.21}$$

where $K \in \mathbb{R}^{(N-1)\times(N-1)}$ is tridiagonal with unit diagonal and off-diagonal entries $-c_n/2$. Lemma A.2 below shows $K \succ 0$, so Eq. (A.21) is strictly negative. Evenness gives a zero derivative at $s = 0$; strict concavity gives the unique maximum at $s = 0$ and the monotonicity on either side. For Eq. (A.16): every weight w_k is positive, every relative error satisfies $1 - R_N(r_k) \ge 1 - R_N(1)$, and $\sum_k w_k = \operatorname{tr} B$; equality requires every generalised eigenvalue to equal one, which is $A = B$. □

A.4.4 The positive tridiagonal form

Lemma A.2 (positive tridiagonal form)**.** *For every integer $N \ge 2$ and every $r > 0$, the matrix K of Eq.* (A.21) *is positive definite.*

Proof. Put $q = \sqrt{r}$ and

$$\omega_n = \frac{n(N-n)}{n + q(N-n)} \ \ (1 \le n \le N-1), \qquad \omega_0 = \omega_N = 0. \tag{A.22}$$

We prove the strict supersolution inequality

$$2\omega_n > c_{n-1}\omega_{n-1} + c_n \omega_{n+1} \qquad (1 \le n \le N-1), \tag{A.23}$$

with the boundary terms omitted (equivalently $c_0 = c_{N-1} = 0$ there). Given Eq. (A.23), for every $z \in \mathbb{R}^{N-1} \setminus \{0\}$,

$$z^\top K z = \sum_{n=1}^{N-1} \frac{(K\omega)_n}{\omega_n} z_n^2 + \frac{1}{2} \sum_{n=1}^{N-2} c_n \omega_n \omega_{n+1} \Big(\frac{z_n}{\omega_n} - \frac{z_{n+1}}{\omega_{n+1}} \Big)^2 > 0, \tag{A.24}$$

since $(K\omega)_n > 0$ by Eq. (A.23); this proves the lemma. It remains to prove Eq. (A.23).

Reduction to one inequality in (a, b, q). Set $a = n$, $b = N - n$ (positive integers) and

$$\begin{aligned} D &= a^2 + b^2 q^2, \qquad D_\pm = (a \pm 1)^2 + (b \mp 1)^2 q^2, \\ Q_- &= a(a-1) + b(b+1)q^2, \qquad Q_+ = a(a+1) + b(b-1)q^2. \end{aligned} \tag{A.25}$$

The left-hand side minus the right-hand side of Eq. (A.23) is

$$S = \frac{2ab}{a + bq} - \frac{D D_-}{Q_-^2} \cdot \frac{(a-1)(b+1)}{a - 1 + (b+1)q} - \frac{D D_+}{Q_+^2} \cdot \frac{(a+1)(b-1)}{a + 1 + (b-1)q}. \tag{A.26}$$

Boundary indices. If $a = b = 1$, $S = 2/(1+q) > 0$. If $a = 1$ and $b \ge 2$, cancellation gives

$$S = \frac{2(b+1)\, H_b(q)}{(bq+1)\big((b-1)q + 2\big)\big(b(b-1)q^2 + 2\big)^2}, \tag{A.27}$$

$$H_b(q) = \Big[b(b-1)q^2 - \frac{b+1}{2} q \Big]^2 + \frac{11b^2 - 18b + 3}{4}\, q^2 + 4, \tag{A.28}$$

and $H_b(q) > 0$ since $11b^2 - 18b + 3 \geq 11$ for $b \geq 2$. The case $b = 1$, $a \geq 2$ follows by exchanging a, b and replacing q by $1/q$, which multiplies S by a positive factor.

Interior indices. For $a, b \geq 2$ set $\alpha = 1/a$, $\beta = 1/b$ and $v = bq/a$, so $0 < \alpha, \beta \leq \frac{1}{2}$, and let

$$H = (a + bq)(a + bq - q + 1)(a + bq + q - 1)\, Q_-^2 Q_+^2 > 0. \tag{A.29}$$

Expanding Eq. (A.26),

$$S = \frac{2q(a+b)^2 a^8}{H}\, \mathcal{P}(v; \alpha, \beta), \qquad \mathcal{P}(v; \alpha, \beta) = v^8\, \mathcal{P}(v^{-1}; \beta, \alpha), \tag{A.30}$$

so it suffices to treat $v \geq 1$. Writing $v = 1 + x$ with $x \geq 0$, $\mathcal{P}(1 + x; \alpha, \beta) = \sum_{j=0}^{8} p_j x^j$ with

$$p_0 = (\alpha + \beta)^2 \big[3(\alpha - \beta)^2 + 4\big], \tag{A.31}$$

$$p_1 = (\alpha + \beta)\big[3\alpha^3 + 21\beta^3 + \alpha(32 - 5\alpha\beta - 19\beta^2)\big], \tag{A.32}$$

$$p_2 = \alpha^4 + \alpha^2(63 - 43\beta^2) + \alpha\beta(38 - 2\alpha^2 + 12\beta^2) + 12 - 49\beta^2 + 64\beta^4, \tag{A.33}$$

$$p_3 = \alpha^2(61 - 45\beta^2) + \alpha\beta(2 + 30\beta^2) + 36 - 131\beta^2 + 111\beta^4, \tag{A.34}$$

$$p_4 = 52 - 168\beta^2 + 120\beta^4 + \alpha^2(32 - 28\beta^2) + \alpha\beta(-32 + 40\beta^2), \tag{A.35}$$

$$p_5 = (1 - \beta^2)\big(44 - 83\beta^2 + 9\alpha^2 - 30\alpha\beta\big), \tag{A.36}$$

$$p_6 = (1 - \beta^2)\big(23 - 36\beta^2 + \alpha^2 - 12\alpha\beta\big), \tag{A.37}$$

$$p_7 = (1 - \beta^2)\big(7 - 9\beta^2 - 2\alpha\beta\big), \tag{A.38}$$

$$p_8 = (1 - \beta^2)^2. \tag{A.39}$$

On $0 < \alpha, \beta \leq \frac{1}{2}$ every coefficient is strictly positive. For p_0, p_1 this is immediate ($32 - 5\alpha\beta - 19\beta^2 \geq 26$). For the others use $\alpha^2, \beta^2, \alpha\beta \leq \frac{1}{4}$, the fact that the β^2-polynomials displayed in p_2, p_3, p_4 are decreasing on $[0, \frac{1}{4}]$, and $-32\alpha\beta \geq -8$ in p_4:

$$p_2 \geq \tfrac{15}{4}, \quad p_3 \geq \tfrac{163}{16}, \quad p_4 \geq \tfrac{19}{2}, \quad p_5 \geq \tfrac{189}{16}, \quad p_6 \geq \tfrac{33}{4}, \quad p_7 \geq \tfrac{51}{16}, \quad p_8 \geq \tfrac{9}{16}. \tag{A.40}$$

Hence $\mathcal{P} > 0$ for $v \geq 1$; the reciprocity in Eq. (A.30) gives $v < 1$; and Eq. (A.23) holds for all indices. The polynomial identities of this proof were also verified symbolically. □

Remark A.3 (scope). Theorem A.1 holds for arbitrary positive-definite Gaussian covariances. For temporal experiments we evaluate the risk on the block-diagonal target $\bar{B}_L$; its relation to trained-model FID is measured separately.

A.5 Stability under approximate matching and the estimator error of TSCM

Theorem A.1 locates the optimum; this subsection shows that it is flat. We write $S_N(A; B)$ for the one-channel risk of Eq. (5) with data covariance B; channels add as in Eq. (5).

Corollary A.1 (stability under approximate matching)**.** *Let $N \geq 2$, $A, B \succ 0$ and $\delta \geq 0$ with $e^{-\delta} B \preceq A \preceq e^{\delta} B$. Then*

$$0 \leq S_N(A; B) - S_N(B; B) \leq \big[R_N(1) - R_N(e^{\delta})\big]\sqrt{\operatorname{tr} B}, \qquad S_N(B; B) = [1 - R_N(1)]\sqrt{\operatorname{tr} B}. \tag{A.41}$$

Moreover $R_N(1) - R_N(e^{\delta}) = c_N \delta^2 + O_N(\delta^4)$ with $c_N = -\frac{1}{2}\frac{d^2}{ds^2} R_N(e^s)\big|_{s=0} > 0$: the excess risk of a near-matched source is second order in the relative spectral error. If $\|B^{-1/2}(A - B)B^{-1/2}\|_{\mathrm{op}} \leq \varepsilon < 1$, the hypothesis holds with $\delta = -\log(1 - \varepsilon)$.

Proof. The r_k of Eq. (5) are the eigenvalues of $B^{-1/2} A B^{-1/2}$, and the two-sided comparison gives $e^{-\delta} \leq r_k \leq e^{\delta}$. By Theorem A.1, R_N is even in $\log r$ and decreasing in $|\log r|$, so $1 - R_N(1) \leq 1 - R_N(r_k) \leq 1 - R_N(e^{\delta})$ for every k. Since $w_k > 0$ and $\sum_k w_k = \operatorname{tr} B$, $S_N(A; B)^2$ lies between $[1 - R_N(1)]^2 \operatorname{tr} B$ and $[1 - R_N(e^{\delta})]^2 \operatorname{tr} B$; take square roots. Commutation of A and B is not used. For the expansion, $s \mapsto R_N(e^s)$ is smooth and even, so its odd derivatives vanish at 0, and $\frac{d^2}{ds^2} R_N(e^s)\big|_{s=0} = R_N(1)\, \frac{d^2}{ds^2} \log R_N(e^s)\big|_{s=0} < 0$ by Eq. (A.21). For the last claim, $\|B^{-1/2}(A - B)B^{-1/2}\|_{\mathrm{op}} \leq \varepsilon$ gives $(1 - \varepsilon)B \preceq A \preceq (1 + \varepsilon)B$ and $1 + \varepsilon \leq (1 - \varepsilon)^{-1}$. □

Numerically, $c_N = 0.0625, 0.0198, 0.0107, 0.0039$ for $N = 2, 4, 8, 24$. At $N = 24$ the floor is $1 - R_{24}(1) = 0.05218$; if every generalised ratio lies in $[0.9, 1.1]$ ($\delta = -\log 0.9$), Eq. (A.41) gives $S_{24}(A; B)/\sqrt{\operatorname{tr} B} \le 0.05222$, an excess of 4.3×10^{-5}. These are values of the exact finite-N product, not of the large-N expansion in Eq. (3); whether TSCM meets such a condition on every mode is an empirical question (Appendix B.7).

Estimation error of TSCM. Per channel and length, the TSCM statistic of Eq. (8) is a shrunk sample second moment of $n = n_L$ windows; under the centred-data assumption, the unshrunk population value is $B_{c,L}$ itself. A fixed estimator family with population limit $A_\infty \neq B$ (such as the Toeplitz variant of Section 3.2, whose centering, lag-0 normalisation, Bartlett taper and zero cross block persist as $n \to \infty$) carries a structural term $\beta > 0$ with $e^{-\beta} B \preceq A_\infty \preceq e^{\beta} B$, and Corollary A.1 applied to the product of the two comparisons bounds its excess divided by $\sqrt{\operatorname{tr} B}$ by $R_N(1) - R_N(e^{\beta+\eta_n})$. This bound tends to the nonzero bias-dependent value $R_N(1) - R_N(e^{\beta})$; it need not equal the actual limiting excess. For centred data, the unshrunk full second moment has no such structural bias; residual means are measured in Table 12, and shrinkage contributes the separate perturbation below. Assume the windows are independent (one per clip), $\mathbb{E}\, xx^\top = B$, and sub-Gaussian relative to B, $\|\langle x, v\rangle\|_{\psi_2} \le K\sqrt{v^\top B v}$ for all v. Then for $u > 0$, with probability at least $1 - 2e^{-u}$ (Vershynin, 2018, Theorem 4.7.1),

$$\left\| B^{-1/2} (\widehat{S}_n - B) B^{-1/2} \right\|_{\mathrm{op}} \le C K^2 \left[\sqrt{\tfrac{L+u}{n}} + \tfrac{L+u}{n} \right] =: \varepsilon_n, \tag{A.42}$$

and the shrinkage moves the estimate by at most $\rho \, \|B^{-1/2}(\mu I - \widehat{S}_n) B^{-1/2}\|_{\mathrm{op}}$, with $\rho \le 0.002$ at every length for the full-data estimates. For the shrunk estimator, put $\widetilde{\varepsilon}_n = \varepsilon_n + \rho \, \|B^{-1/2}(\mu I - \widehat{S}_n) B^{-1/2}\|_{\mathrm{op}}$. If $\widetilde{\varepsilon}_n < 1$, use $\eta_n = -\log(1 - \widetilde{\varepsilon}_n)$ in Corollary A.1. The excess risk divided by $\sqrt{\operatorname{tr} B}$ is at most $R_N(1) - R_N(e^{\eta_n}) = c_N \eta_n^2 + O(\eta_n^4)$. The sampling contribution is of order $K^4(L + u)/n$ in the small-error regime; the same rate holds for the shrunk estimator if its relative shrinkage perturbation is of the same order as ε_n. A small shrinkage intensity alone does not ensure this near small eigenvalues. On HumanML3D, L/n is 0.003 at $L = 64$ ($n = 22{,}498$) and 0.023 at $L = 196$ ($n = 8{,}620$). The plain relative operator-norm errors of subsampled estimates against the full one stay below $\sqrt{L/n}$ at every budget (Appendix B.7); the whitened error of Eq. (A.42) does not, because the near-floor eigenvalues of the full estimate are far below the shrinkage floor of a small-sample one, so the corollary's bound is loose there and the risk is measured directly. The independent-reference excess of the full estimate is 0.3% of the floor. For this full-data estimate, the relative spectral error $\delta = \max_k |\log r_k|$ has median 1.3 over channels against the independent reference: it is set by the smallest eigenvalues, near the 10^{-3} noise floor, which carry almost no risk weight w_k. The corollary is therefore stated with δ but used with the risk itself.

A.6 Cost of the block-diagonal surrogate

Section 2.2 proves the general Gaussian result; TSCM applies it to $\mathcal{N}(0, \bar{B})$, retaining within-channel temporal covariances and dropping the cross-channel blocks of B. Requiring a channel-independent source does not by itself justify this model choice: for a target with cross-channel covariance, the channel-independent source that minimizes Euler endpoint error need not be $\bar{B}$. We measure the difference on a Gaussian target. For $\mathcal{N}(0, B)$ and a block-diagonal source A, $N = 24$ uniform Euler steps on the covariance-driven field G_t of Eq. (1) map x_0 to Mx_0 with $M = \prod_n (I + G_{t_n}/N)$, and we report the exact W_2 distance between $\mathcal{N}(0, MAM^\top)$ and $\mathcal{N}(0, B)$. When A and B commute this equals $S_N(A; B)$ of Eq. (5); otherwise S_N is only an upper bound, so the exact distance is used throughout. We compare white noise, TSCM ($A = \bar{B}$), the best channel-independent source (the better of a per-channel rescaling $a_c \bar{B}_c$ and a free block-diagonal SPD matrix, both optimized by L-BFGS from $\bar{B}$), and $A = B$, which is not channel independent and gives the floor.

The target has four channels of $L = 16$ frames with Ornstein–Uhlenbeck kernels of length $1, 2, 4, 8$ and a common-driver cross-channel correlation ρ: $B_{cc} = K_c$ and $B_{cd} = \rho \, K_c^{1/2} K_d^{1/2}$, so $\bar{B} = \operatorname{diag}(K_1, \ldots, K_4)$ for every ρ. Table 4 gives the excess W_2. The excess of TSCM over the best channel-independent source grows with ρ but stays below the price of channel independence itself and far below the excess of white noise, and the best channel-independent source is $\bar{B}$ scaled up by a common factor: the surrogate gets the temporal shape right and only the overall scale slightly cold.

In the extreme two-channel, one-frame case $B = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$, where $\bar{B} = I$, the best isotropic source is likewise hotter than I; at $\rho = 0.99$ it beats I by 5.5% for $N = 2$ and by 1.8% for $N = 24$. This is oracle risk on a Gaussian target, not FID.

Table 4: Excess endpoint W_2 ($N = 24$) on the synthetic target as a function of the cross-channel correlation ρ, and the scale of the best per-channel rescaling of $\bar{B}$.

ρ	0	0.2	0.4	0.6	0.8	0.9	0.95
TSCM over best channel-independent	0.0%	0.0%	0.4%	1.6%	3.8%	5.4%	6.6%
best channel-independent over $A = B$	0.0%	0.9%	3.0%	5.9%	8.8%	9.8%	9.9%
white over TSCM	14%	15%	17%	19%	22%	24%	25%
best scale a_c of $\bar{B}_c$	1.00	1.06	1.22	1.48	1.85	2.10	2.29

A.7 Explicit Runge–Kutta bounds; grids, paths and couplings

A.7.1 A source-dependent resolution scale

For one mode put

$$\tau_r = \frac{r}{1+r}, \qquad \delta_r = \frac{\sqrt{r}}{1+r} = \frac{1}{2\Lambda(r)}, \qquad d_r(t) = \sqrt{(t-\tau_r)^2 + \delta_r^2}. \tag{A.43}$$

Then the gain is the real part of a simple pole,

$$g(t;r) = \frac{t-\tau_r}{(t-\tau_r)^2 + \delta_r^2} = \operatorname{Re} \frac{1}{t - \tau_r + i\delta_r}, \tag{A.44}$$

and consequently, for every integer $j \geq 0$,

$$\left|\partial_t^j g(t;r)\right| \leq \frac{j!}{d_r(t)^{j+1}} \leq j! \left[2\Lambda(r)\right]^{j+1}. \tag{A.45}$$

The exact solution is a multiple of $\sqrt{D(t)}$, and its relative acceleration is

$$\frac{z''(t)}{z(t)} = g'(t;r) + g(t;r)^2 = \frac{\delta_r^2}{d_r(t)^4}, \qquad \sup_{t\in[0,1]} \left|\frac{z''(t)}{z(t)}\right| = 4\Lambda(r)^2. \tag{A.46}$$

Thus Λ controls the time derivatives of the field as well as its gain. It is a resolution scale of the covariance-driven field; it is not by itself a statement that the full learned ODE is stiff.

A.7.2 A one-step and endpoint theorem for explicit Runge–Kutta methods

Let $\mathcal{M}$ be a fixed finite-stage explicit Runge–Kutta method of classical order $p \geq 1$ for non-autonomous ODEs (standard tableau formulation, see Hairer & Zennaro, 1996). Let $\mathcal{T} = (0 = t_0 < \cdots < t_N = 1)$ be a grid with $h_n = t_{n+1} - t_n$ and $h_{\max} = \max_n h_n$; write M_n for the multiplier of $\mathcal{M}$ on $z' = g(t;r)z$ over step n and

$$E_n = \sqrt{\frac{D(t_{n+1})}{D(t_n)}} > 0 \tag{A.47}$$

for the exact multiplier.

Theorem A.2 (source-dependent error on a non-uniform grid)**.** *There exist constants $\eta_{\mathcal{M}}, C_{\mathcal{M}}, K_{\mathcal{M}} > 0$ depending only on the method such that, if*

$$\frac{h_n}{d_r(t_n)} \leq \eta_{\mathcal{M}} \qquad \textit{for every } n, \tag{A.48}$$

then $M_n > 0$ and

$$\left|\frac{M_n}{E_n} - 1\right| \leq C_{\mathcal{M}} \Big(\frac{h_n}{d_r(t_n)}\Big)^{p+1}. \tag{A.49}$$

The normalised endpoint amplitude $R_{\mathcal{M},\mathcal{T}}(r) = \sqrt{r}\prod_n M_n$ satisfies

$$\big|R_{\mathcal{M},\mathcal{T}}(r) - 1\big| \le \exp\Big\{C_{\mathcal{M}} \sum_n \Big(\frac{h_n}{d_r(t_n)}\Big)^{p+1}\Big\} - 1 \tag{A.50}$$

$$\le \exp\big\{K_{\mathcal{M}}\big[2\Lambda(r)h_{\max}\big]^p\big\} - 1. \tag{A.51}$$

In particular, for fixed r the endpoint error is $O_{\mathcal{M}}\big((\Lambda(r)h_{\max})^p\big)$ in the resolved regime.

Proof. At a step starting at t put $d = d_r(t)$, $a = (t-\tau_r)/d \in [-1,1]$ and the local time $\xi = (u-t)/d$. The ODE becomes

$$\frac{dz}{d\xi} = F_a(\xi)\, z, \qquad F_a(\xi) = \frac{a+\xi}{1+2a\xi+\xi^2}, \tag{A.52}$$

whose exact multiplier from 0 to $\eta = h/d$ is $E_a(\eta) = \sqrt{1+2a\eta+\eta^2}$. Choose $\eta_0 > 0$ so that all stage times satisfy $|c_i\eta| < \frac{1}{2}$ for $|\eta| \le \eta_0$; the denominators in Eq. (A.52) are then uniformly separated from zero for $a \in [-1,1]$. On this compact parameter set the finite explicit stage recursion and E_a have bounded derivatives through order $p+1$; classical order p makes their Taylor coefficients agree through degree p; and Taylor's theorem together with $E_a(\eta) \ge 1 - |\eta|$ gives $|M_a/E_a - 1| \le C_{\mathcal{M}}|\eta|^{p+1}$ uniformly in a. Decreasing η_0 makes the right-hand side less than $\frac{1}{2}$, which gives positivity. This proves Eq. (A.49) with $\eta_{\mathcal{M}} = \eta_0$.

Since $\sqrt{r}\prod_n E_n = 1$, write $M_n = E_n(1+\epsilon_n)$ with $|\epsilon_n|$ bounded by Eq. (A.49); then

$$|R_{\mathcal{M},\mathcal{T}} - 1| \le \prod_n (1+|\epsilon_n|) - 1 \le \exp\Big(\sum_n |\epsilon_n|\Big) - 1, \tag{A.53}$$

which is Eq. (A.50). For $u \in [t_n, t_{n+1}]$ the distance function d_r is 1-Lipschitz, so $d_r(u) \le d_r(t_n) + h_n \le (1+\eta_{\mathcal{M}})d_r(t_n)$. Therefore

$$\sum_n \frac{h_n^{p+1}}{d_r(t_n)^{p+1}} \le h_{\max}^p\,(1+\eta_{\mathcal{M}})^{p+1} \int_0^1 d_r(t)^{-p-1}\,dt \tag{A.54}$$

$$\le h_{\max}^p\,(1+\eta_{\mathcal{M}})^{p+1}\,\delta_r^{-p}\,I_p, \qquad I_p = \int_{\mathbb{R}} (1+s^2)^{-(p+1)/2}\,ds < \infty. \tag{A.55}$$

Substituting $\delta_r^{-1} = 2\Lambda(r)$ and absorbing $(1+\eta_{\mathcal{M}})^{p+1}I_p$ into $K_{\mathcal{M}}$ proves Eq. (A.51). □

Remark A.4 (what the theorem does and does not say). The theorem covers explicit Heun and classical RK4 and does not require a uniform grid. It is a uniform upper bound on the exact covariance-driven field, not a monotonicity statement for each method's realised endpoint error. A Taylor expansion of the Heun multiplier gives

$$M_n - E_n = \frac{h_n^3}{12}\big(g'' - 2g^3\big)(t_n) + O(h_n^4) = -\frac{h_n^3}{2}\,\frac{(t_n-\tau_r)\,\delta_r^2}{d_r(t_n)^6} + O(h_n^4), \tag{A.56}$$

whose sign changes across τ_r, so higher-order cancellation must not be replaced by a universal lower bound. With a single Heun step, $g(0) = -1$ makes the predictor zero, so $M = \frac{1}{2}$ and $R = \sqrt{r}/2$: the error vanishes at $r = 4$, not at $r = 1$. The unique optimum of Theorem A.1 is therefore specific to uniform Euler. The local argument applies pathwise to an adaptively selected grid whose accepted steps satisfy Eq. (A.48), but a sample-dependent grid need not induce a linear Gaussian endpoint map, so exact Gaussian-oracle interpretations should use deterministic grids. Predictor–corrector and exponential diffusion solvers such as UniPC (Zhao et al., 2023) are not covered by an explicit-RK statement; their update formulas require a separate analysis.

A.7.3 Optimal grids for a single Euler mode

Proposition A.2 (equal-angle Euler grids)**.** *For any increasing N-step grid $\mathcal{T}$, the exact covariance-field Euler amplitude of one mode is*

$$R_{\mathcal{T}}(r) = \prod_{n=0}^{N-1} \cos\theta_n, \qquad \theta_n = \phi_r(t_{n+1}) - \phi_r(t_n), \qquad \phi_r(t) = \arctan\frac{t}{(1-t)\sqrt{r}}. \tag{A.57}$$

For $N \geq 2$,

$$\max_{\mathcal{T}} R_{\mathcal{T}}(r) = \cos^N\Big(\frac{\pi}{2N}\Big), \tag{A.58}$$

attained uniquely by equal angular increments, i.e. by the grid

$$t_n = \frac{\sqrt{r}\,\tan\big(n\pi/(2N)\big)}{1+\sqrt{r}\,\tan\big(n\pi/(2N)\big)} \quad (0 \leq n < N), \qquad t_N = 1. \tag{A.59}$$

Proof. The telescoping cosine identity in the proof of Lemma A.1 holds for every step length h_n, because $P(t) = ((1-t)\sqrt{r}, t)$ is affine in t. The angles sum to $\pi/2$. Strict concavity of $\log\cos$ on $(0, \pi/2)$ and Jensen's inequality give the bound and the equality condition; inverting ϕ_r gives the grid. □

Remark A.5 (source–grid co-design). The maximum in Eq. (A.58) does not depend on r: an individually adapted grid removes the mismatch-dependent part of the one-mode Euler error. Different ratios, however, want different grids, whereas a multivariate sampler shares one grid across all modes. If $A = \kappa B$, all ratios agree and a common equal-angle grid serves every mode, so a free source scale is not uniquely determined when source and grid are optimised jointly. This is source–grid co-design, not a universal fixed-source lower bound.

A.7.4 Smooth interpolation schedules

For independent endpoints and $x_t = \alpha(t)x_0 + \beta(t)x_1$, the covariance-driven coefficient is

$$G_t = \big[\alpha\dot{\alpha}A + \beta\dot{\beta}B\big]\big[\alpha^2 A + \beta^2 B\big]^{-1}, \tag{A.60}$$

diagonal in the same generalised eigenbasis, with mode-wise gain

$$g_{\alpha,\beta}(t; r) = \frac{\alpha\dot{\alpha}\, r + \beta\dot{\beta}}{\alpha^2 r + \beta^2}. \tag{A.61}$$

With $c = \alpha + \beta > 0$ and $s = \beta/c$, exactly

$$g_{\alpha,\beta}(t; r) = \frac{\dot{c}}{c} + \dot{s}\, g(s; r). \tag{A.62}$$

When s increases from 0 to 1, normalising the state by $c(t)$ and using s as time reduces the mode to the linear-path family, and Theorem A.2 applies on the induced s-grid. A solver applied directly in the original state and time coordinates is not invariant under this transformation, and its error constants also depend on the schedule derivatives: for $\alpha, \beta \in C^{p+1}$ with bounded derivatives, c bounded away from zero and $0 \leq s \leq 1$, differentiating Eq. (A.62) and using Eq. (A.45) bounds $|\partial_t^j g_{\alpha,\beta}|$ by $C_{\mathrm{path},j}[1+\Lambda(r)]^{j+1}$ for $0 \leq j \leq p$. Even the discretisation floor depends on the path: for $\alpha(t) = \cos(\pi t/2)$, $\beta(t) = \sin(\pi t/2)$ and $r = 1$, $g_{\alpha,\beta} \equiv 0$ and there is no covariance-field integration error. The linear-path floor should not be carried unchanged to another schedule.

A.7.5 Coupled endpoints and the optimal-transport boundary

Let $K = \mathrm{Cov}(x_0, x_1)$ with centred endpoints. Then, whenever Σ_t is invertible,

$$\Sigma_t = \alpha^2 A + \beta^2 B + \alpha\beta\,(K + K^\top), \tag{A.63}$$

$$C_t = \dot{\alpha}\alpha A + \dot{\beta}\beta B + \dot{\alpha}\beta K + \dot{\beta}\alpha K^\top, \tag{A.64}$$

$$G_t = C_t \Sigma_t^{-1}, \tag{A.65}$$

which are covariance and linear-regression identities; for jointly Gaussian endpoints the population field is exactly linear, otherwise its regression residual remains. For a scalar linear path with source/data ratio r and endpoint correlation ρ, let

$$D_{r,\rho}(t) = (1-t)^2 r + t^2 + 2\rho\sqrt{r}\, t(1-t). \tag{A.66}$$

Whenever $D_{r,\rho} > 0$ the Gaussian field has $g_{r,\rho} = D'_{r,\rho}/(2D_{r,\rho})$ and

$$g'_{r,\rho} + g_{r,\rho}^2 = \frac{r(1-\rho^2)}{D_{r,\rho}(t)^2}. \tag{A.67}$$

Trajectory curvature is therefore set jointly by the mismatch and the coupling. The monotone scalar Gaussian OT coupling has $\rho = 1$,

$$x_1 = \frac{x_0}{\sqrt{r}}, \qquad z(t) = \frac{(1-t)\sqrt{r} + t}{\sqrt{r}}\, z(0), \tag{A.68}$$

an affine trajectory on which Euler is exact on every grid for every $r > 0$; straight trajectories are also the explicit aim of OT-flow constructions (Kornilov et al., 2024). A mismatch-only error lower bound therefore cannot hold across arbitrary couplings; a coupling-aware extension must retain K or a curvature quantity such as Eq. (A.67). The empirical TSCM construction uses independent coupling, $\rho = 0$; the next subsection extends the theory to an aligned correlated family.

A.8 Covariance shape, scalar calibration, and endpoint cancellation

The resolution bound in Appendix A.7 does not identify an optimizer for a higher-order solver. We now separate two questions: which covariance *shape* can be optimal, and which scalar *scale* the path and solver select. We then show that matching has an endpoint cancellation property for even-order Runge–Kutta methods.

A.8.1 A Gaussian family with aligned endpoint correlations

Let $A, B \succ 0$ on $\mathbb{R}^d$ and write

$$Q = B^{-1/2} A B^{-1/2} = V \operatorname{diag}(r_k) V^\top, \qquad z = V^\top B^{-1/2} x.$$

In this subsection we extend the independent-coupling setting to centred jointly Gaussian endpoints with

$$\operatorname{Cov}(z_0) = \operatorname{diag}(r_k), \quad \operatorname{Cov}(z_1) = I, \quad \operatorname{Cov}(z_0, z_1) = \rho \operatorname{diag}(\sqrt{r_k}), \quad -1 < \rho < 1. \tag{A.69}$$

The scalar correlation ρ is common to all modes and held fixed as A varies; the original-coordinate cross covariance is not held fixed. This is an aligned family, not an arbitrary coupling. Independent coupling is $\rho = 0$.

For a smooth scalar path $z_t = \alpha(t) z_0 + \beta(t) z_1$, assume $(\alpha(0), \beta(0)) = (1, 0)$ and $(\alpha(1), \beta(1)) = (0, 1)$. Where the path covariance is positive, the exact population field separates into

$$z_k' = g(t; r_k) z_k, \qquad g(t; r) = \frac{\partial_t D(t; r)}{2 D(t; r)}, \qquad D(t; r) = \alpha(t)^2 r + \beta(t)^2 + 2\rho\sqrt{r}\, \alpha(t)\beta(t). \tag{A.70}$$

In particular, the exact endpoint multiplier is $r^{-1/2}$.

A.8.2 An actual Wasserstein optimum with the target covariance shape

Consider a fixed deterministic path, grid, and numerical method whose update acts separately on these linear modes, with the same continuous signed normalized response $R(r)$ for every mode. Explicit Runge–Kutta methods have this property. Other methods require checking their update formulas. Let $I \subset (0, \infty)$ be a connected admissible interval on which the responses are defined, set $a(r) = |R(r)|$, and assume

$$e_* := \min_{r \in I} |a(r) - 1|$$

is attained at some $\kappa_* \in I$. Denote the numerical endpoint law for source covariance A and its corresponding population field by P^A.

Theorem A.3 (shape matching with scalar calibration)**.** *Under these assumptions,*

$$\min_{\substack{A \succ 0 \\ \operatorname{spec}(B^{-1/2} A B^{-1/2}) \subset I}} W_2\big(P^A, \mathcal{N}(0, B)\big) = e_* \sqrt{\operatorname{tr} B}, \qquad A_* = \kappa_* B \text{ attains the minimum.} \tag{A.71}$$

No commutation assumption on admissible A and B is needed.

Proof. The endpoint covariance is

$$C(A) = B^{1/2} V \operatorname{diag}\big(a(r_k)^2\big) V^\top B^{1/2}. \tag{A.72}$$

If $e_* = 0$, $A = \kappa_* B$ gives $C(A) = B$. If $e_* > 0$, continuity and connectedness imply that either $a(r) \le 1 - e_*$ for all $r \in I$, or $a(r) \ge 1 + e_*$ for all $r \in I$: taking values on both sides of one would give a zero by the intermediate value theorem. Thus either $C(A) \preceq (1 - e_*)^2 B$ or $C(A) \succeq (1 + e_*)^2 B$.

For any coupling of centred random vectors X, Z, the reverse triangle inequality in L^2 gives

$$\left(\mathbb{E}\|X - Z\|^2\right)^{1/2} \ge \left|\sqrt{\operatorname{tr} \operatorname{Cov}(X)} - \sqrt{\operatorname{tr} \operatorname{Cov}(Z)}\right|. \tag{A.73}$$

Taking the infimum over couplings proves $W_2 \ge e_* \sqrt{\operatorname{tr} B}$. At $A = \kappa_* B$, $C(A) = a(\kappa_*)^2 B$, and the scaling coupling attains this lower bound. □

Corollary A.2 (uniform Euler also minimizes actual endpoint W_2)**.** *For independent coupling, a linear path, and $N \ge 2$ uniform Euler steps, $A = B$ uniquely minimizes the actual Gaussian endpoint W_2, including when the competing covariances do not commute with B. The minimum is $[1 - R_N(1)]\sqrt{\operatorname{tr} B}$.*

Proof. Theorem A.1 gives $0 < R_N(r) \le R_N(1) < 1$, with equality only at $r = 1$. Apply Theorem A.3 with $I = (0, \infty)$. If any $r_k \ne 1$, Eq. (A.72) has strictly smaller trace than $R_N(1)^2 B$, since $v_k^\top B v_k > 0$. The trace lower bound is then strictly larger than the matched error. □

Theorem A.3 asserts the existence of an optimum with the target covariance shape. Multiple scalar minimizers may also yield non-proportional optima. The assumption that the scalar minimum is attained is essential on a noncompact interval; for example, it does not follow from continuity alone. Each candidate uses its own population field, so the theorem does not authorize changing the source of a frozen trained model.

A.8.3 Endpoint cancellation for even-order Runge–Kutta methods

Theorem A.4 (reflection-induced endpoint superconvergence)**.** *In Eq.* (A.70)*, let $\alpha, \beta \in C^{p+3}$ on a neighbourhood of $[0, 1]$, let $\alpha(t) = \beta(1 - t)$, and suppose $D(t; 1)$ is bounded away from zero there. Fix an explicit Runge–Kutta method of even classical order $p \ge 2$. On deterministic reflection-symmetric grids,*

$$t_{N-n} = 1 - t_n, \qquad h_{\max} = \max_n (t_{n+1} - t_n),$$

its normalized endpoint response satisfies

$$R_{\mathcal{M},\mathcal{T}}(1) - 1 = O(h_{\max}^{p+1}). \tag{A.74}$$

Uniformly for $|s| \le \delta_0$ in a sufficiently small fixed neighbourhood and sufficiently fine grids,

$$|R_{\mathcal{M},\mathcal{T}}(e^s) - 1| \le C\big(|s| h_{\max}^p + h_{\max}^{p+1}\big). \tag{A.75}$$

Consequently, if $\|\log Q\|_{\mathrm{op}} \le \delta \le \delta_0$, the Gaussian endpoint error obeys

$$W_2\big(P^A, \mathcal{N}(0, B)\big) \le C\sqrt{\operatorname{tr} B}\,\big(\delta h_{\max}^p + h_{\max}^{p+1}\big). \tag{A.76}$$

Constants depend on the method, path, fixed correlation, and spectral neighbourhood, not on the number or minimum size of the steps.

Proof. At $r = 1$, $D(1-t; 1) = D(t; 1)$ and hence $g(1-t; 1) = -g(t; 1)$. Write $M(t, h)$ and $E(t, h)$ for numerical and exact scalar multipliers. For sufficiently small signed steps both are positive, and order p and smoothness give, uniformly in t,

$$\ell(t, h) := \log \frac{M(t, h)}{E(t, h)} = a(t) h^{p+1} + O(h^{p+2}), \qquad a \in C^1.$$

Reflecting every Runge–Kutta stage sends (t, h) to $(1-t, -h)$. The two sign changes, in the field and in the step size, cancel in each stage recursion, so $M(1-t, -h) = M(t, h)$. The exact multiplier has the same property. Comparison of the leading coefficients yields $a(1 - t) = (-1)^{p+1} a(t) = -a(t)$.

Let $h_n = t_{n+1} - t_n$. The mirrored interval starts at $1 - t_n - h_n$ and has the same length. Pairing intervals gives

$$\sum_n h_n^{p+1} a(t_n) = \frac{1}{2} \sum_n h_n^{p+1} \big[a(t_n) + a(1 - t_n - h_n)\big]$$
$$= \frac{1}{2} \sum_n h_n^{p+1} \big[a(t_n) - a(t_n + h_n)\big] = O\Big(\sum_n h_n^{p+2}\Big) = O(h_{\max}^{p+1}).$$

The local remainders have the same summed bound. Exact multipliers telescope to one at $r = 1$, which proves Eq. (A.74) after exponentiating.

For $r = e^s$ near one, smooth dependence gives $a(t; s) = a(t; 0) + O(|s|)$ uniformly, and exact normalized multipliers still telescope to one. The additional sum is bounded by $C|s| \sum_n h_n^{p+1} \le C|s| h_{\max}^p$. This proves Eq. (A.75). The mode-wise Gaussian pairing used in Proposition 3, whose weights sum to $\operatorname{tr} B$, gives Eq. (A.76). □

The Runge–Kutta method itself need not be symmetric: explicit Heun and classical RK4 have respectively $O(N^{-3})$ and $O(N^{-5})$ endpoint errors at the match on uniform grids. The cancellation concerns the endpoint, not the whole trajectory. A nonsymmetric path or grid need not retain it. The theorem also does not assert that the matched scale exactly minimizes every finite-step error.

A.8.4 Heun: an explicit leading term and a calibrated scale

Proposition A.3 (Heun asymptotics)**.** *For the linear path with the aligned correlation $-1 < \rho < 1$, uniform Heun has*

$$R_{\mathrm{H},N}(r, \rho) = 1 + \frac{1 - \rho^2}{8N^2}(r - r^{-1}) + O(N^{-3}). \tag{A.77}$$

For each fixed ρ, $r = 1$ is the unique fixed ratio that cancels the N^{-2} coefficient. Under independent coupling,

$$R_{\mathrm{H},N}(1) = 1 - \frac{3(4 + \pi)}{32N^3} + O(N^{-4}), \qquad r_N = 1 + \frac{3(4 + \pi)}{8N} + O(N^{-2}), \tag{A.78}$$

where, for all sufficiently large N, r_N is the unique solution of $R_{\mathrm{H},N}(r_N) = 1$ in a fixed neighbourhood of one.

Proof. Set $k = g' + g^2 = r(1 - \rho^2)/D^2$ using Eq. (A.67). Expansion of the Heun multiplier gives the leading local logarithmic error

$$a_3(t) = \frac{g'' - 2g^3}{12} = \frac{1}{8} k'(t).$$

With $h = 1/N$, summing gives

$$\log R_{\mathrm{H},N}(r, \rho) = h^2 \int_0^1 a_3(t)\, dt + O(h^3) = \frac{h^2}{8}(1 - \rho^2)(r - r^{-1}) + O(h^3),$$

which proves Eq. (A.77). To obtain the next term, the local h^4 coefficient is

$$a_4 = \frac{g'''}{24} - \frac{(g')^2}{8} - \frac{g^2 g'}{4} + \frac{g^4}{8}.$$

Euler–Maclaurin summation of the $h^3 a_3$ term subtracts $\frac{1}{2}\int_0^1 a_3'$, and $a_4 - a_3'/2 = (g^4 - (g')^2)/8$. Therefore, for independent coupling,

$$\log R_{\mathrm{H},N}(r) = \frac{r - r^{-1}}{8N^2} + \frac{1}{8N^3} \int_0^1 \big[g^4 - (g')^2\big] dt + O(N^{-4}). \tag{A.79}$$

At $r = 1$, putting $u = 2t - 1$ reduces the integral coefficient to

$$\frac{1}{8} \int_0^1 \big[g^4 - (g')^2\big] dt = \int_{-1}^1 \frac{2u^2 - 1}{(1 + u^2)^4}\, du = -\frac{3(4 + \pi)}{32}.$$

The expansion and its first r derivative are uniform near one. The leading function $(r - r^{-1})/8$ is strictly increasing there, with derivative $1/4$ at one. Thus $N^2 \log R_{\mathrm{H},N}(r)$ has a unique zero in that neighbourhood for all large N; expansion at one gives r_N in Eq. (A.78). □

For comparison, one Heun step in the correlated family gives exactly

$$R_{\mathrm{H},1}(r,\rho) = \rho + \frac{1-\rho^2}{2}\sqrt{r}, \qquad \kappa_* = \frac{4}{(1+\rho)^2}. \tag{A.80}$$

Indeed, $g(0) = -1 + \rho/\sqrt{r}$ and $g(1) = 1 - \rho\sqrt{r}$; substitution into the two-stage update proves the formula. It recovers the one-step value $\kappa_* = 4$ at $\rho = 0$. At $N = 24$ the independent-coupling root is approximately 1.11830945. These are exact-field calibrations, not trained-model hyperparameter recommendations.

A.8.5 Joint optimization of covariance and a shared Euler grid

Theorem A.5 (a common optimal grid requires matched covariance shape)**.** *In the linear-path Gaussian family of Eq.* (A.69)*, fix $0 \le \rho < 1$ and $N \ge 2$. Put $\theta = \arccos\rho$ and $c_{N,\rho} = \cos^N(\theta/N)$. Over all positive-definite sources and common deterministic N-step grids,*

$$\min_{A,\mathcal{T}} W_2\big(P^A_{\mathrm{E},\mathcal{T}}, \mathcal{N}(0,B)\big) = (1 - c_{N,\rho})\sqrt{\operatorname{tr} B}. \tag{A.81}$$

Equality holds exactly when $A = \kappa B$ for some $\kappa > 0$ and

$$t_n = \frac{\sqrt{\kappa}\sin(n\theta/N)}{\sin((N-n)\theta/N) + \sqrt{\kappa}\sin(n\theta/N)}, \qquad 0 \le n \le N. \tag{A.82}$$

Proof. For a single mode consider

$$P_r(t) = \big((1-t)\sqrt{r} + \rho t,\ t\sqrt{1-\rho^2}\big).$$

Its squared norm is $D(t;r,\rho)$ and its polar angle increases from zero to $\theta \le \pi/2$. The telescoping argument of Lemma A.1 applies to this affine line on any grid, giving

$$R_{\mathcal{T}}(r,\rho) = \prod_n \cos\Delta\phi_n, \qquad \sum_n \Delta\phi_n = \theta.$$

Every angular increment is in $(0, \pi/2)$. Strict concavity of $\log\cos$ gives $0 < R_{\mathcal{T}}(r,\rho) \le c_{N,\rho}$, with equality exactly at equal angles. Thus $C(A) \preceq c^2_{N,\rho} B$, and Eq. (A.73) gives the stated lower bound. Inverting the polar angle gives Eq. (A.82), which attains it when every ratio equals κ.

If any mode fails to reach $c_{N,\rho}$, the trace inequality is strict and the lower bound cannot be attained. Therefore every mode must have equal angular increments. Its first interior node is

$$t_1 = \frac{\sqrt{r_k}\sin(\theta/N)}{\sin((N-1)\theta/N) + \sqrt{r_k}\sin(\theta/N)},$$

which is strictly increasing in $\sqrt{r_k}$. A common t_1 forces all ratios to agree, and the remaining nodes are then fixed by equal angles. □

Thus a shared grid can absorb a common scale, while matched covariance shape lets all modes use that grid optimally. The restriction $\rho \ge 0$ ensures positive cosine factors on every admissible grid. At $\rho = 1$ the aligned coupling is deterministic and the trajectories are affine: Euler is exact for every A and grid, and the shape necessity disappears.

A.9 Proportional covariances and the transport distance

If $A = \kappa B$ then $r_k = \kappa$ for every k, $G_t = g(t;\kappa)I$, and all modes reach source/data variance balance at the same time $t = \sqrt{\kappa}/(1+\sqrt{\kappa})$; but $\Lambda = \cosh(\frac{1}{2}\log\kappa)$, so synchronisation does not make the field mild. Conversely, the Bures distance

$$W_2^2(A,B) = \operatorname{tr} A + \operatorname{tr} B - 2\operatorname{tr}\left(A^{1/2}BA^{1/2}\right)^{1/2} \tag{A.83}$$

is $\sum_k(\sqrt{a_k} - \sqrt{b_k})^2$ in the commuting case: two sources with the same W_2 can put their mismatch into a mode with small b_k (large r_k, harmless in W_2 but $\Lambda \to \infty$) or spread it evenly, so equal W_2 is compatible with arbitrarily different Λ and S_N. A direct comparison with real HumanML3D windows likewise shows that empirical source–data distances need not order few-step FID (Appendix B.6, Table 10).

A.10 Positive definiteness and the level of a Toeplitz autocorrelation

For a one-channel window x of length L, write $B_L = \mathrm{Cov}(x)$, $u = \mathbf{1}/\sqrt{L}$, $H = I - uu^\top$, and $W_L = \mathrm{Cov}(Hx) = HB_LH$. The full level–fluctuation decomposition is

$$\begin{aligned} B_L &= H\,W_L\,H + d_L\,uu^\top + (\text{level–fluctuation cross terms}),\\ d_L &= u^\top B_L u = L\cdot \mathrm{Var}\big(\text{mean of } L \text{ frames}\big). \end{aligned} \tag{A.84}$$

The cross terms are $HB_Luu^\top + uu^\top B_LH$. For a Toeplitz matrix $A = \mathrm{Toep}(\rho)$ with $\rho_0 = 1$, expanding $u^\top Au$ gives

$$u^\top Au = 1 + 2\sum_{h=1}^{L-1}\Big(1 - \frac{h}{L}\Big)\rho_h. \tag{A.85}$$

In Eq. (9), $\hat{d}^{\mathrm{F}}_{c,L}$ is this level computed from the uncentered training ACF. The following lemma gives the positive-definiteness guarantee for the centered-ACF construction.

Lemma A.3. *Let $h^{(n)} \in \mathbb{R}^{L_n}$ be the centred training sequences of one channel and*

$$\rho_\ell = \frac{\sum_n \sum_i h_i^{(n)} h_{i+\ell}^{(n)}}{\sum_n \sum_i \big(h_i^{(n)}\big)^2} \tag{A.86}$$

the Bartlett autocorrelation. Then $\mathrm{Toep}(\rho)$ is positive definite for every window length, and its level is Eq. (A.85)*.*

Proof. Extending each sequence by zeros, ρ is the Fourier coefficient sequence of $S(\omega) \propto \sum_n \big|\sum_j h_j^{(n)} e^{-ij\omega}\big|^2 \geq 0$, so for any $c \neq 0$,

$$c^\top \mathrm{Toep}(\rho)\, c = \frac{1}{2\pi}\int \Big|\sum_i c_i e^{ii\omega}\Big|^2 S(\omega)\, d\omega > 0, \tag{A.87}$$

both factors being nonzero trigonometric polynomials. Eq. (A.85) is $u^\top \mathrm{Toep}(\rho) u$ written out. □

Two consequences are used in Section 3.2. For white data ($\rho_\ell = 0$ for $\ell \geq 1$) the level of the Toeplitz matrix is 1, while the level of the centred autocorrelation of finite windows has expectation tending to $\sigma^2/3$. For data with a persistent per-sequence offset the level $d_L = L\,\mathrm{Var}(\text{mean})$ grows like L, whereas the Bartlett estimate of ρ_ℓ decays with ℓ (only the $L_n - \ell$ pairs of each sequence contribute to lag ℓ, against a fixed denominator), so Eq. (A.85) grows more slowly than d_L (55 against 94 at $L = 196$). Centering a sequence removes its mean, so the centred autocorrelation estimates the fluctuation block HW_LH of Eq. (A.84) and carries no information about the level; the residual level $14 \to 7$ of the centred Toeplitz source is a window artefact, not an estimate.

A.11 Conditional flows and the marginal covariance

Throughout this subsection B denotes the full covariance on the state space; a channel-wise surrogate may be formed afterwards by taking diagonal channel blocks. Conditional centering and temporal sequence centering are different operations: Hx (within-sequence fluctuation) is not $x - \mu_Y$ (within-condition randomness), and both conditional components enter both temporal projections,

$$HBH = HWH + HVH, \qquad u^\top Bu = u^\top Wu + u^\top Vu. \tag{A.88}$$

Put

$$\mu_y = \mathbb{E}[x_1 \mid Y = y],\;\; B_y = \mathrm{Cov}(x_1 \mid Y = y),\;\; \mu = \mathbb{E}[x_1],\;\; W = \mathbb{E}_Y[B_Y],\;\; V = \mathrm{Cov}_Y(\mu_Y). \tag{A.89}$$

The law of total covariance then reads

$$B = W + V. \tag{A.90}$$

Assume $x_0 \sim \mathcal{N}(0, A)$ independent of (x_1, Y) and $x_t = (1-t)x_0 + tx_1$.

A.11.1 WHICH COVARIANCE DESCRIBES WHICH FIELD

Proposition A.4 (conditional and shared linear projections)**.** *For each y, whenever the displayed covariance is invertible,*

$$v_{\mathrm{c}}(x,t,y) = \mu_y + G_{y,t}\,(x - t\mu_y) + \rho_{y,t}(x), \qquad G_{y,t} = \big[tB_y - (1-t)A\big]\big[(1-t)^2A + t^2B_y\big]^{-1}, \tag{A.91}$$

where the residual has zero conditional mean and is conditionally orthogonal to $x_t - t\mu_y$; it vanishes if $x_1 \mid y$ is Gaussian. Separately, the best condition-independent affine predictor of the conditional population field under the joint interpolation law is

$$\mu + G_t\,(x_t - t\mu), \qquad G_t = \big[tB - (1-t)A\big]\big[(1-t)^2A + t^2B\big]^{-1}, \tag{A.92}$$

with the marginal coefficient of Eq. (1)*.*

Proof. Conditionally on $Y = y$, the means of x_t and $x_1 - x_0$ are $t\mu_y$ and μ_y, the state covariance is $(1-t)^2A + t^2B_y$, and their cross covariance is $tB_y - (1-t)A$. Linear regression and the tower property give Eq. (A.91); the Gaussian assertion is Gaussian conditional expectation. For the shared projection let $Z_t = x_t - t\mu$. The tower property gives

$$\mathbb{E}\big[(v_{\mathrm{c}}(x_t,t,Y) - \mu)Z_t^\top\big] = \mathbb{E}\big[(x_1 - x_0 - \mu)Z_t^\top\big] = tB - (1-t)A, \tag{A.93}$$

and $\mathrm{Cov}(Z_t) = (1-t)^2A + t^2B$. Together these give Eq. (A.92) and the joint orthogonality of its residual. □

Remark A.6 (how guidance relates to the shared projection). With $v_{\mathrm{u}}(x,t) = \mathbb{E}[v_{\mathrm{c}}(x_t,t,Y) \mid x_t = x]$ and $\Delta = v_{\mathrm{c}} - v_{\mathrm{u}}$,

$$v_w(x,t,y) = \mu + G_t(x - t\mu) + \rho_{u,t}(x) + w\,\Delta(x,t,y), \qquad \mathbb{E}\big[\Delta(x_t,t,Y) \mid x_t\big] = 0, \tag{A.94}$$

so the shared projection is unchanged for every deterministic $w(t)$ under the original interpolation law. Guided generation changes the joint law of state and condition, so this orthogonality is not an endpoint covariance guarantee. The marginal oracle of Section 2 is the precise justification for a marginal-covariance shared proxy; it does not identify B with every B_y, and it does not turn the proxy into a bound on conditional-model numerical error.

A.11.2 A COUNTEREXAMPLE TO UNIVERSAL MARGINAL MATCHING

Example A.1. Let $Y \sim \mathcal{N}(0,3)$ and $\epsilon \sim \mathcal{N}(0,1)$ be independent and $x_1 = Y + \epsilon$. Then $B = 4$, $B_y = 1$, and the marginal target is itself Gaussian. On the exact conditional field, Euler preserves the mean t_nY exactly on any grid; subtracting that mean reduces the recursion to the centred covariance field with target variance 1. With two uniform steps,

$$A = 1: \quad W_2 = 1 - R_2(1) = \tfrac{1}{2}, \qquad\qquad A = B = 4: \quad W_2 = 1 - R_2(4) = \tfrac{3}{5}. \tag{A.95}$$

The marginal-matched source is strictly worse for conditional covariance integration. For every uniform step count $N \geq 2$, Theorem A.1 makes $A = 1 = B_y$ optimal for this conditional family. Marginal matching is therefore not universally optimal for conditional integration; the following bound quantifies its cost relative to the best shared source.

This counterexample rules out universal condition-wise optimality for marginal matching, which TSCM does not claim. TSCM uses pooled marginal temporal statistics without a conditional-moment model; the following results quantify the cost of that shared-source choice and how additional conditional information can change its optimum.

A.11.3 AVERAGE CONDITIONAL ERROR WITH A SINGLE SHARED SOURCE

We now compare marginal matching with the best *shared* source, rather than requiring it to optimize every condition separately. Assume

$$x_1 \mid Y = y \sim \mathcal{N}(\mu_y, B_y), \qquad x_0 \sim \mathcal{N}(0,A), \qquad x_0 \perp (x_1, Y),$$

use a linear path and $N \geq 2$ uniform Euler steps on the exact conditional field ($w = 1$), and let $P^y_{N,A}$ be the conditional endpoint law. The marginal target may be a non-Gaussian mixture. Define

$$\mathcal{R}_N(A) := \mathbb{E}_Y W_2^2\big(P^Y_{N,A}, \mathcal{N}(\mu_Y, B_Y)\big), \qquad \varepsilon_N := 1 - R_N(1). \tag{A.96}$$

This compares output and target at the same condition. It is not the Wasserstein distance after mixing over conditions, or a trained-model FID. The conditional field in Eq. (A.91) has zero residual here. Its Euler mean obeys $m_{n+1} = m_n + h\mu_y$ when $m_n = t_n\mu_y$, so all mean errors vanish exactly and the numerical error is in the covariance.

Lemma A.4 (a global scalar excess-error envelope)**.** *For every $N \geq 2$, the constant*

$$\kappa_N := \sup_{\substack{r>0 \\ r\neq 1}} \frac{[1 - R_N(r)]^2 - \varepsilon_N^2}{r + r^{-1} - 2} \tag{A.97}$$

satisfies $0 < \kappa_N < \infty$ and $\kappa_N = O(N^{-2})$, with an absolute implied constant. In particular, $\kappa_2 = 1/16$.

Proof. Write $s = \log r$. Corollary A.1 gives $R_N(e^s) = R_N(1) - c_N s^2 + O(s^4)$ with $c_N > 0$, so the ratio in Eq. (A.97) tends to $2\varepsilon_N c_N > 0$ at $s = 0$. At either infinite endpoint its numerator is bounded and its denominator diverges. Continuity therefore proves positivity and finiteness.

To obtain a bound uniform in N, use the angle increments of Lemma A.1. With $\phi_r(t) = \arctan[t/((1-t)\sqrt{r})]$, $\phi_r' = \sqrt{r}/D(t)$ and $\theta_n = \int_{t_n}^{t_{n+1}} \phi_r'(t)\,dt$. Product subadditivity, $1 - \cos\theta \leq \theta^2/2$, and Cauchy–Schwarz on each interval give

$$1 - R_N(r) \leq \frac{1}{2}\sum_n \theta_n^2 \leq \frac{1}{2N}\int_0^1 \frac{r}{D(t)^2}\,dt = \frac{1 + \frac{\pi}{2}\Lambda(r)}{2N}. \tag{A.98}$$

For $|s| \geq 1$, $r + r^{-1} - 2 = 4(\Lambda(r)^2 - 1)$, and Eq. (A.98) bounds the ratio by a constant times N^{-2}.

For $|s| \leq 1$, $\phi_{e^s}'(t)$ and its first two s derivatives are uniformly bounded. Thus θ_n and its first two s derivatives are $O(N^{-1})$. For all sufficiently large N, differentiating $\log R_N(e^s) = \sum_n \log\cos\theta_n(s)$ twice gives uniform $O(N^{-1})$ bounds on this function and its first two derivatives. The same orders hold for $1 - R_N$ and its derivatives. Hence $\partial_s^2[1 - R_N(e^s)]^2 = O(N^{-2})$. Evenness and Taylor's theorem show that the numerator of Eq. (A.97) is $O(N^{-2}s^2)$, while its denominator is at least s^2. The finitely many remaining N are absorbed into the constant.

For $N = 2$, put $\lambda = \Lambda(r) \geq 1$. Since $R_2(r) = 1/(2\lambda)$, the ratio equals

$$\frac{3\lambda - 1}{16\lambda^2(\lambda + 1)}.$$

It is strictly decreasing for $\lambda \geq 1$ and has limit $1/16$ at one. □

To avoid confusing covariance heterogeneity with the temporal projector H, write

$$\mathcal{H} := \mathbb{E}_Y[(B_Y - W)^2], \qquad \mathcal{D}(A) := \mathbb{E}_Y \operatorname{tr}\left[(A - B_Y)A^{-1}(A - B_Y)\right]. \tag{A.99}$$

Both $\mathcal{H} \succeq 0$ and $\mathcal{D}(A) \geq 0$; $\mathcal{D}(A)$ equals $\mathbb{E}\|A^{-1/2}(A - B_Y)\|_F^2$.

Theorem A.6 (average conditional error and marginal-source regret)**.** *Assume $B = W + V \succ 0$ and $\mathbb{E}\|B_Y\|_F^2 < \infty$. For every shared $A \succ 0$,*

$$\varepsilon_N^2 \operatorname{tr} W \leq \mathcal{R}_N(A) \leq \varepsilon_N^2 \operatorname{tr} W + \kappa_N \mathcal{D}(A). \tag{A.100}$$

In particular,

$$0 \leq \mathcal{R}_N(B) - \inf_{A\succ 0} \mathcal{R}_N(A) \leq \kappa_N\left[\operatorname{tr}(VB^{-1}V) + \operatorname{tr}(B^{-1}\mathcal{H})\right]. \tag{A.101}$$

The infimum compares condition-independent shared sources. These statements require neither commuting covariances nor pointwise spectral comparability of B_Y with B, and remain valid for positive-semidefinite B_Y.

Proof. First fix a condition with covariance $C \succ 0$ and set $Q_C = C^{-1/2}AC^{-1/2}$. Functional calculus and Lemma A.4 give

$$\varepsilon_N^2 I \preceq [I - R_N(Q_C)]^2 \preceq \varepsilon_N^2 I + \kappa_N(Q_C + Q_C^{-1} - 2I).$$

The Gaussian mode pairing bounds the actual squared Wasserstein distance by $\operatorname{tr}(C[I-R_N(Q_C)]^2)$. Therefore

$$W_2^2(P^y_{N,A}, \mathcal{N}(\mu_y, C)) \le \varepsilon_N^2 \operatorname{tr} C + \kappa_N \big[\operatorname{tr} A + \operatorname{tr}(CA^{-1}C) - 2 \operatorname{tr} C \big] \\ = \varepsilon_N^2 \operatorname{tr} C + \kappa_N \operatorname{tr} \big[(A-C)A^{-1}(A-C) \big]. \tag{A.102}$$

For the lower bound, Theorem A.1 implies

$$C_{\text{out}} = C^{1/2} R_N(Q_C)^2 C^{1/2} \preceq R_N(1)^2 C.$$

The trace lower bound of Eq. (A.73) consequently yields $W_2^2 \ge \varepsilon_N^2 \operatorname{tr} C$. This is a lower bound on actual W_2, not only on the mode-pairing cost.

For singular C, replace it by $C + \eta I$ and let $\eta \downarrow 0$. Every Euler evaluation has $t_n < 1$, so $(1 - t_n)^2 A + t_n^2 C \succ 0$ even at $\eta = 0$. Its update and Gaussian endpoint covariance therefore depend continuously on C. Continuity of the Gaussian Wasserstein distance proves the same two bounds at every $C \succeq 0$. Averaging these pointwise bounds gives Eq. (A.100); the moment assumption makes its upper bound finite.

Finally write $B - B_Y = V - (B_Y - W)$. Since $\mathbb{E}(B_Y - W) = 0$, the cross terms vanish and trace cyclicity gives

$$\mathcal{D}(B) = \operatorname{tr}(VB^{-1}V) + \mathbb{E} \operatorname{tr} \big[(B_Y - W) B^{-1} (B_Y - W) \big] = \operatorname{tr}(VB^{-1}V) + \operatorname{tr}(B^{-1}\mathcal{H}).$$

Subtract the common lower bound on $\inf_A \mathcal{R}_N(A)$ to obtain Eq. (A.101). □

Only the shared covariance is inverted in Eq. (A.101). Small conditional eigenvalues therefore do not automatically invalidate the bound. If $B_y = 0$, the exact conditional field is $(\mu_y - x)/(1 - t)$, and Euler on any grid gives $x_{n+1} - \mu_y = [(1 - t_{n+1})/(1 - t_n)](x_n - \mu_y)$; its final step reaches μ_y exactly. Large variance ratios alone do not imply large absolute conditional endpoint error.

In Example A.1, $W = 1$, $V = 3$, $B = 4$ and $\mathcal{H} = 0$. At $N = 2$, Eq. (A.101) gives a regret bound of $(1/16)(9/4) = 9/64 = 0.140625$, whereas the exact regret is $(3/5)^2 - (1/2)^2 = 0.11$. The counterexample is thus included in the quantitative guarantee.

For $\operatorname{tr} W > 0$, one also has

$$\frac{\mathcal{R}_N(B)}{\inf_{A \succ 0} \mathcal{R}_N(A)} \le 1 + \frac{\kappa_N}{\varepsilon_N^2} \frac{\operatorname{tr}(VB^{-1}V) + \operatorname{tr}(B^{-1}\mathcal{H})}{\operatorname{tr} W}. \tag{A.103}$$

Absolute regret decreases as $O(N^{-2})$ for a fixed finite budget; relative near-optimality requires the normalized budget to be small. Appendix A.11.7 estimates this budget at caption-cluster resolutions on HumanML3D. Marginal covariance statistics alone do not establish it. These results concern $w = 1$: the CFG bias of Proposition 4 is a separate change in the continuous-time flow, even though both involve V.

A.11.4 The shared covariance that minimizes the upper bound

Proposition A.5 (an optimal shared upper bound)**.** *Let $M = \mathbb{E}[B_Y^2] = W^2 + \mathcal{H} \succ 0$. The unique minimizer of $\mathcal{D}(A)$ over $A \succ 0$ is*

$$A_{\text{bound}} = M^{1/2} = (W^2 + \mathcal{H})^{1/2}, \qquad \min_A \mathcal{D}(A) = 2 \operatorname{tr}(M^{1/2} - W). \tag{A.104}$$

For a homoscedastic family $B_y = W \succ 0$ almost surely, $A = W$ uniquely minimizes the actual risk $\mathcal{R}_N$, with value $\varepsilon_N^2 \operatorname{tr} W$, regardless of the variation in conditional means.

Proof. Expansion and trace cyclicity give

$$\mathcal{D}(A) = \operatorname{tr} A - 2 \operatorname{tr} W + \operatorname{tr}(A^{-1}M).$$

For any $A \succ 0$,

$$\operatorname{tr} A + \operatorname{tr}(A^{-1}M) - 2 \operatorname{tr} M^{1/2} = \|A^{1/2} - A^{-1/2}M^{1/2}\|_F^2 \ge 0,$$

with equality exactly when $A = M^{1/2}$. In the homoscedastic case this is $A = W$, and every conditional covariance is simultaneously matched. Corollary A.2 proves attainment and uniqueness for the actual risk. □

For heterogeneous conditional covariances, A_{bound} minimizes the proved upper bound; it need not minimize $\mathcal{R}_N$ itself. If M is singular, the same infimum is approached by $A = M^{1/2} + \eta I$ but is not attained by a positive-definite matrix with that singular limit.

Nor is W generally optimal for the actual shared-source risk. In one dimension, let $\mu_Y = 0$ and let B_Y be 1 or 4 with equal probability, so $W = 5/2$. At $N = 2$, $R_2(r) = \sqrt{r}/(1+r)$ gives $\mathcal{R}_2(a) = \frac{1}{2}\sum_{b\in\{1,4\}} b[1 - R_2(a/b)]^2$. Direct evaluation yields $\mathcal{R}_2(3) \approx 0.67105 < \mathcal{R}_2(5/2) \approx 0.67764$.

An offline conditional-mean correction. For a deterministic predictor $m(y)$, or a predictor fitted independently of the evaluation pair (x_1, Y), conditioning on Y gives the exact identity

$$\mathbb{E}[(x_1 - m(Y))(x_1 - m(Y))^\top] = W + \mathbb{E}[(\mu_Y - m(Y))(\mu_Y - m(Y))^\top]. \tag{A.105}$$

Thus $m \equiv 0$ gives B in the centred setting, whereas perfect mean estimation gives W. The estimation error remains as an explicit positive-semidefinite term. This suggests an offline residual source that is still condition-independent at sampling time, but is not the TSCM construction evaluated here; it would require consistent source use during training and sampling. It also does not estimate M or remove covariance heterogeneity by itself. One real example per exact condition is insufficient to identify B_y without additional assumptions; covariance within text clusters is not automatically covariance at a fixed condition.

A.11.5 Why a conditional heterogeneity budget is necessary

Proposition A.6 (no uniform conditional rate from the marginal covariance alone)**.** *For every fixed scalar source variance $a > 0$ and every finite $N \geq 2$,*

$$\sup_{\substack{x_1|Y \text{ Gaussian} \\ \mu_Y = 0,\ \mathbb{E}B_Y = 1}} \mathcal{R}_N(a) = 1. \tag{A.106}$$

The supremum is unchanged if every conditional variance is positive and every individual family has a finite second moment of B_Y.

Proof. For any such scalar family,

$$\mathcal{R}_N(a) = \mathbb{E}\big[B_Y(1 - R_N(a/B_Y))^2\big] \leq \mathbb{E}B_Y = 1.$$

Fix $0 < \eta < 1$, and let $B_Y = \eta$ with probability $1 - p$ and $B_Y = [1 - (1-p)\eta]/p$ with probability p. The mean variance is one and both variances are positive. As $p \downarrow 0$, the rare condition has ratio $a/B_Y \to 0$, so its contribution to $\mathcal{R}_N(a)$ tends to $1 - \eta$ by Lemma A.1. Thus the supremum is at least $1 - \eta$ for every η, proving equality. Every fixed two-point family has finite $\mathbb{E}B_Y^2$, but this moment, and hence $\mathcal{H}$, diverges as $p \downarrow 0$. □

For each fixed distribution, dominated convergence still gives $\mathcal{R}_N(a) \to 0$ as $N \to \infty$. What fails is a rate uniform over all conditional families with the same marginal second moment. The budget in Eq. (A.101) supplies the additional information, without requiring a lower spectral bound on every B_y.

A.11.6 A separate worst-mode transfer under spectral comparability

Proposition A.7 (transfer of a marginal spectral match)**.** *Suppose, for some $\epsilon, \gamma \geq 0$,*

$$e^{-\epsilon}B \preceq A \preceq e^{\epsilon}B, \qquad e^{-\gamma}B \preceq B_y \preceq e^{\gamma}B \quad \textit{for the conditions considered.} \tag{A.107}$$

Then every conditional generalised variance ratio lies in

$$e^{-(\epsilon+\gamma)} \leq r_{y,k} \leq e^{\epsilon+\gamma}, \qquad \textit{and hence} \qquad \Lambda(r_{y,k}) \leq \cosh\frac{\epsilon+\gamma}{2}. \tag{A.108}$$

Proof. Combining the two Loewner inequalities gives $e^{-(\epsilon+\gamma)}B_y \preceq A \preceq e^{\epsilon+\gamma}B_y$. Congruence by $B_y^{-1/2}$ and the Rayleigh quotient bound the eigenvalues of $B_y^{-1/2}AB_y^{-1/2}$; the formula for Λ proves the second claim. □

This transfers the bounds of Theorem A.2 to the conditional fields, subject to their step-size condition and, for non-Gaussian conditionals, the same residual limitation. Spectral comparability is an additional assumption: it does not follow from Eq. (A.90), and conditioning may make a covariance nearly singular. Appendix A.11.8 measures channel-wise γ on regularised second-moment surrogates at a HumanML3D caption-cluster resolution, using data held out from the source estimate, together with the Euler risk of a training-derived source against each cluster surrogate.

A.11.7 THE CONDITIONAL BUDGET ON HUMANML3D

We evaluate the right-hand side of Eq. (A.101) at a fixed caption-cluster resolution $Z = f(Y)$, replacing Y by Z throughout. Clustering supplies repeated observations for a measurable diagnostic; $\mathrm{Cov}(x_1 \mid Z)$ need not equal the covariance at a fixed caption $Y = y$. With $W_Y = \mathbb{E}\,\mathrm{Cov}(x_1 \mid Y)$ and $W_Z = \mathbb{E}\,\mathrm{Cov}(x_1 \mid Z)$, total covariance gives $W_Z = W_Y + \mathbb{E}\,\mathrm{Cov}(\mu_Y \mid Z) \succeq W_Y$. Thus the cluster budget concerns a coarsened conditional Gaussian surrogate, not the full-text conditional regret; coarsening does not make its normalized budget a conservative bound for Y. We use k-means clustering in a 64-dimensional truncated-SVD space of caption tf–idf vectors. The measurement uses $L = 64$, $K \in \{8, 16, 32\}$ caption clusters, 2,605 motion identities and 5,642 windows from the validation and test splits, with five redraws; budget terms are normalised by $\tau = \mathrm{tr}\, W_Z$.

Numerical scalar constant. An extended-precision (`longdouble`) dense search of the ratio in Eq. (A.97) over $10^{-3} \le |\log r| \le 60$, combined with its analytic $r \to 1$ limit $2\varepsilon_N c_N$, gives numerical estimates $\widehat{\kappa}_{24} = 4.057460 \times 10^{-4}$ and $\widehat{\kappa}_{16} = 8.8073 \times 10^{-4}$: no larger value is found on the search grid, and the scalar envelope is numerically verified there. The limit avoids cancellation near $r = 1$. This equality is not assumed for every step count: at $N = 3$ the search finds a larger ratio, approximately 1.7850×10^{-2} near $r = 5.005$ and its reciprocal, compared with the limit 1.6128×10^{-2}. These are numerical checks of the supremum; Lemma A.4 supplies the global existence and $O(N^{-2})$ guarantee.

Debiased budget estimate. Using $\widehat{\kappa}_{24}$ and debiased estimates of the budget terms (with a four-group correction for the mean-variation term), the estimated right-hand side $\widehat{\kappa}_{24}[\mathrm{tr}(VB^{-1}V) + \mathrm{tr}(B^{-1}\mathcal{H})]/\tau$ is 0.00023, 0.00031 and 0.00035 for $K = 8, 16, 32$, respectively, or 8.5%, 11.5% and 12.9% of the floor $\varepsilon_{24}^2 = 0.00272$. These are numerical estimates of the budget at resolution Z, relative to that Gaussian squared-error floor; $\inf_A \mathcal{R}_{24}(A)$ is not measured.

Empirical-family check. Taking the half-sample cluster covariances C_k (with a 10^{-3} noise floor) as a finite conditional Gaussian family, we use exactly the same C_k and cluster weights in $\mathcal{R}_{24}$, $\mathcal{D}$ and τ; with the training-set TSCM source and the other-half W and B as the three shared-source choices, both sides of Eq. (A.100) hold numerically in all $3 \times 2 \times 2 \times 3 = 36$ combinations of cluster count, caption/random labels, half-sample and source, and the tightness ratio $(\mathcal{R}_{24} - \varepsilon_{24}^2\tau)/(\widehat{\kappa}_{24}\mathcal{D})$ ranges from 0.27 to 0.89 (0.65–0.89 for TSCM). This check concerns the plug-in empirical family; its risk is compared with its own budget, not with the debiased budget estimate above.

A.11.8 HELD-OUT CAPTION-CLUSTER SECOND-MOMENT TRANSFER

We measure spectral variation of regularised caption-cluster second moments as a surrogate diagnostic for Proposition A.7 at the resolutions $K \in \{8, 16, 32\}$ of Appendix A.11.7, using the held-out validation and test splits, and score a fixed training-derived source against each cluster. It does not test spectral comparability for the full text condition and uses no trained model or generated samples.

Protocol and spectral summaries. We use the 5,642 standardised 64-frame windows and five redraws of Appendix A.11.7. For channel c, let $B_c^{(2)}$ be the pooled diagnostic split's temporal second moment and $C_{k,c}^{(2)}$ the second moment of cluster k, each regularised with the noise floor $10^{-3}\,\mathrm{tr}(\widehat{S})/L$ of Appendix B.6. Here $\widehat{S}$ is the corresponding unregularised second moment and $L = 64$. Before regularisation, with $\mu_{k,c} = \mathbb{E}[x_c \mid Z = k]$, $\mathbb{E}[x_c x_c^\top \mid Z = k] = \mathrm{Cov}(x_c \mid Z = k) + \mu_{k,c}\mu_{k,c}^\top$. No per-window temporal mean or cluster mean is removed. We use these second moments as covariance matrices of centred Gaussian surrogates; they are not covariances at a fixed

text condition. The reference $B_c^{(2)}$ is held-out, while the evaluated TSCM source is estimated on training data.

For each cluster and channel, diagonalise $Q_{k,c} = (B_c^{(2)})^{-1/2} C_{k,c}^{(2)} (B_c^{(2)})^{-1/2} = V_{k,c}\,\mathrm{diag}(r_{k,c,j}) V_{k,c}^\top$ and define

$$\gamma_{k,c}^{\max} = \max_j |\log r_{k,c,j}|, \qquad \gamma_{k,c}^{90} = \mathrm{quantile}_{0.9,j}\,|\log r_{k,c,j}|,$$
$$M_{k,c} = \left(\frac{\sum_j w_{k,c,j} \log^2 r_{k,c,j}}{\sum_j w_{k,c,j}} \right)^{1/2}, \quad w_{k,c,j} = v_{k,c,j}^\top C_{k,c}^{(2)} v_{k,c,j}. \tag{A.109}$$

We first compute each statistic within one channel, then take its median over 263 channels and average this median over five window redraws; M_k denotes this per-cluster average for the weighted RMS. Thus $\gamma_{k,c}^{\max}$ is a channel-specific Loewner radius, whereas its reported median describes a typical channel, not a bound uniform over all channels. The spectral RMS uses the displayed target-second-moment weights. Oracle risks are evaluated separately with the target-whitened eigenpairs and weights of Eq. (5). Random labels reassign the same motion identities with the same cluster sizes and provide a finite-sample control alongside each caption result.

Table 5: **Caption-cluster second-moment shift and surrogate Euler risk on held-out HumanML3D.** For the three spectral statistics, entries are medians over clusters of the per-channel median averaged over five redraws; brackets give ranges over clusters. Random labels use the same identities and cluster sizes. The last two columns are cluster medians of the relative TSCM excess e_k and risk ratio q_k defined below. The last row compares training and held-out marginals. All matrices are regularised second-moment surrogates.

K	labels	$\gamma^{\max}$	γ^{90}	M_k	TSCM excess	TSCM/ white
8	caption	2.06 [1.69, 2.51]	1.37 [0.98, 1.81]	0.78 [0.55, 1.05]	6.3%	0.39
	random	1.77 [1.34, 2.10]	1.00 [0.70, 1.30]	0.43 [0.26, 0.57]	2.9%	0.38
16	caption	2.45 [1.75, 3.29]	1.74 [1.02, 2.48]	0.96 [0.65, 1.19]	9.7%	0.41
	random	2.13 [1.50, 2.66]	1.30 [0.83, 1.79]	0.57 [0.32, 0.88]	4.2%	0.39
32	caption	2.83 [2.18, 4.92]	1.99 [1.38, 3.48]	1.05 [0.78, 2.18]	9.8%	0.41
	random	2.46 [2.06, 3.92]	1.56 [1.26, 2.59]	0.76 [0.55, 1.29]	6.8%	0.39
train vs. held-out		1.11	0.57	0.167	0.36%	0.38

How far the caption-cluster second moments move. Table 5 gives caption-cluster medians of $\gamma^{\max}$ from 2.06 to 2.83, compared with weighted RMS M_k from 0.78 to 1.05. The worst-mode summary is conservative near the regularisation floor, as in the estimator analysis of Appendix B.7; random labels already yield substantial spectral spread, with $M_k = 0.43$–0.76. Caption-cluster variation exceeds that control at every resolution. The training-to-held-out marginal comparison gives much smaller summaries, 1.11, 0.57 and 0.167, respectively: split-level second-moment transfer is substantially more stable than variation across caption clusters.

Surrogate Euler cost for a shared source. In each redraw, let $C_k^{(2)} = \mathrm{diag}(C_{k,c}^{(2)})$ and score the training TSCM table A_{TSCM}, white noise and the cluster-matched source using $s_k(A) = S_{24}(A; C_k^{(2)})/\sqrt{\mathrm{tr}\, C_k^{(2)}}$. Let $\bar{s}_k(A)$ denote its average over the five redraws. The matched floor is $\varepsilon_{24} = 1 - R_{24}(1) \approx 0.05218$ in every redraw. The table reports $e_k = \bar{s}_k(A_{\mathrm{TSCM}})/\varepsilon_{24} - 1$ and $q_k = \bar{s}_k(A_{\mathrm{TSCM}})/\bar{s}_k(I)$. Across all 56 caption clusters, the TSCM excess has median 9.5%, and 46 clusters are within 20% of their own matched floor; the white-noise median excess at each resolution is approximately 169–176%. The TSCM/white risk ratio has median 0.41 and range 0.18–0.61. TSCM has lower oracle risk in every cluster: the smallest normalised gain $\bar{s}_k(I) - \bar{s}_k(A_{\mathrm{TSCM}})$ is $0.84\varepsilon_{24}$. For the pooled held-out reference, its excess is only 0.36%, consistent with Appendix B.7.

Figure 6 uses the same target-weighted, per-channel shift statistic as Table 5. Larger shift is associated with greater TSCM excess within each resolution (Pearson correlations 0.67–0.85), while the

advantage over white noise stays positive throughout the observed range and shows no consistent decreasing trend across resolutions.

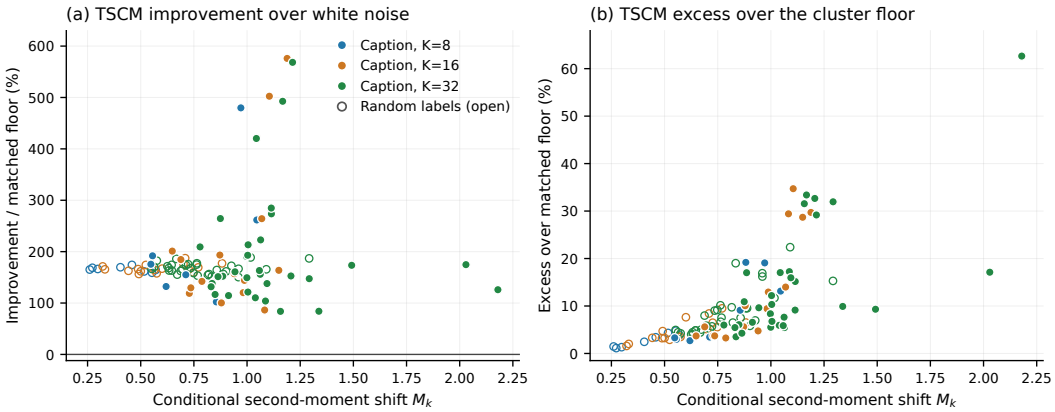

Figure 6: **Caption-cluster second-moment shift and the numerical cost of a shared source.** Each filled point is a caption cluster at $K = 8, 16, 32$; open points show the random-label control. The horizontal axis is the target-weighted, per-channel-median M_k of Table 5, averaged over five redraws. **Left**: improvement over white noise, $100[\bar{s}_k(I) - \bar{s}_k(A_{\mathrm{TSCM}})]/\varepsilon_{24}$, positive for every caption cluster. **Right**: TSCM excess over the cluster-matched floor, $100e_k$.

These measurements support transfer of the source's numerical advantage across held-out caption-cluster subpopulations: the training-derived source lowers the covariance-only Euler risk in every measured caption-cluster surrogate. The spectral summaries and source scores are distinct from the budget decomposition $\mathcal{D}(A)$ of Appendix A.11.7. They provide evidence for the numerical premise of Appendix A.15, without evaluating semantic OOD generation.

A.12 A SHARP LOCAL COVARIANCE GUARANTEE BEYOND GAUSSIAN TARGETS

The endpoint results above concern Gaussian population fields. A different objective gives a sharp result for arbitrary target distributions: the covariance deficit of one Euler step *started from the prescribed interpolation law*. Throughout this subsection, $x_0 \sim \mathcal{N}(0, A)$ is independent of (x_1, Y), $\mathbb{E}x_1 = 0$, $\mathrm{Cov}(x_1) = B \succ 0$, and $A \succ 0$, except in the explicitly correlated extension below. No Gaussian or pointwise spectral assumption is made on $x_1 \mid Y$.

A.12.1 AN EXACT ONE-STEP IDENTITY AND THE EFFECT OF CONDITIONING

Put $U = x_1 - x_0$, $x_t = (1-t)x_0 + tx_1$, and

$$v_{\mathrm{u}} = \mathbb{E}[U \mid x_t], \qquad v_{\mathrm{c}} = \mathbb{E}[U \mid x_t, Y], \qquad \Delta = v_{\mathrm{c}} - v_{\mathrm{u}}.$$

Write

$$\Xi_t^{\mathrm{u}} = \mathbb{E}\,\mathrm{Cov}(U \mid x_t), \qquad \Xi_t^{\mathrm{c}} = \mathbb{E}\,\mathrm{Cov}(U \mid x_t, Y).$$

Proposition A.8 (Euler covariance deficit)**.** *For $0 \le t < t + h \le 1$, one conditional Euler step $\widehat{x}_{t+h} = x_t + hv_{\mathrm{c}}$ satisfies*

$$\mathrm{Cov}(x_{t+h}) - \mathrm{Cov}(\widehat{x}_{t+h}) = h^2\Xi_t^{\mathrm{c}}, \qquad \Xi_t^{\mathrm{c}} = \Xi_t^{\mathrm{u}} - \mathbb{E}[\Delta\Delta^\top] \preceq \Xi_t^{\mathrm{u}}. \tag{A.110}$$

Let $L_t = tB - (1-t)A$, $\Sigma_t = (1-t)^2A + t^2B$, and

$$\Gamma_t = A + B - L_t\Sigma_t^{-1}L_t^\top. \tag{A.111}$$

Then

$$0 \preceq \Xi_t^{\mathrm{c}} \preceq \Xi_t^{\mathrm{u}} \preceq \Gamma_t. \tag{A.112}$$

The last inequality is an equality for a Gaussian target; the entire chain is attained when, in addition, Y is independent of x_1.

Proof. Since $x_{t+h} = x_t + hU$ and $\widehat{x}_{t+h} = \mathbb{E}[x_{t+h} \mid x_t, Y]$, the law of total covariance gives the first identity. Applying the same law to U with nested information $\sigma(x_t) \subset \sigma(x_t, Y)$ gives the second: $\mathbb{E}[\Delta \mid x_t] = 0$.

The best linear predictor is $G_t x_t = L_t \Sigma_t^{-1} x_t$, with residual covariance Γ_t. Orthogonality of conditional expectation gives the matrix decomposition

$$\Gamma_t = \Xi_t^{\mathrm{u}} + \mathbb{E}\big[(v_{\mathrm{u}} - G_t x_t)(v_{\mathrm{u}} - G_t x_t)^\top\big].$$

This proves the envelope. Joint Gaussianity makes the conditional expectation linear, and an independent label supplies no additional information. □

The first identity and the linear-prediction inequality only require finite second moments and the appropriate invertible state covariance; Gaussianity of the source is not needed for these identities themselves. Conditional information reduces this local covariance deficit. This statement is for the population conditional field ($w = 1$); a CFG field is generally not a conditional expectation and need not satisfy it.

A.12.2 Minimax calibration with only marginal second moments

Define the integrated normalized local deficit

$$\mathcal{J}(A; P_{x_1,Y}) := \int_0^1 \mathrm{tr}(\Sigma_t^{-1} \Xi_t^{\mathrm{c}})\, dt. \tag{A.113}$$

It measures the one-step covariance loss per unit h^2, normalized by the current covariance and integrated over time. Let $\mathcal{P}(B)$ contain all joint target–condition laws with $\mathbb{E}x_1 = 0$ and $\mathrm{Cov}(x_1) = B$; no condition information or conditional covariance is fixed.

Theorem A.7 (sharp minimax local calibration)**.** *Let r_k be the eigenvalues of $Q = B^{-1/2} A B^{-1/2}$ on $\mathbb{R}^d$. Then*

$$\mathcal{J}_*(A; B) := \sup_{P \in \mathcal{P}(B)} \mathcal{J}(A; P) = d + \frac{\pi}{2} \sum_{k=1}^{d} \cosh\left(\frac{1}{2} \log r_k\right). \tag{A.114}$$

The supremum is attained by $x_1 \sim \mathcal{N}(0, B)$ with an independent condition. Its unique minimizer over $A \succ 0$ is $A = B$, and globally

$$\mathcal{J}_*(A; B) - \mathcal{J}_*(B; B) \geq \frac{\pi}{16} \| \log(B^{-1/2} A B^{-1/2}) \|_F^2. \tag{A.115}$$

The same value and optimum hold for unconditional targets, by omitting Y.

Proof. Proposition A.8 supplies the bound by $\int \mathrm{tr}(\Sigma_t^{-1} \Gamma_t) dt$, attained for all t simultaneously by the stated Gaussian target and independent label. The trace is invariant under an invertible linear change of coordinates. In the generalized mode coordinates, Σ_t is diagonal with entries $D(t; r_k)$, and the linear-prediction error covariance has entries $r_k / D(t; r_k)$. Consequently

$$\int_0^1 \mathrm{tr}(\Sigma_t^{-1} \Gamma_t)\, dt = \sum_k \int_0^1 \frac{r_k}{[(1-t)^2 r_k + t^2]^2}\, dt = \sum_k \left[1 + \frac{\pi}{2} \Lambda(r_k)\right],$$

using the integral from Appendix A.3. Each summand is uniquely minimized at $r_k = 1$. Finally $\cosh(s/2) - 1 \geq s^2/8$ gives Eq. (A.115). □

For channel-independent source design, the theorem applies to each channel and the local budgets can be added, fixing the within-channel target covariances. This does not assert that a block-diagonal source optimizes an unrestricted objective using the full cross-channel covariance. More fundamentally, Eq. (A.110) starts every step from the true interpolation law. Numerical trajectories leave that law, so the identity cannot be iterated into a non-Gaussian endpoint bound without further control. The Gaussian field is a sharp worst case for this *local* objective, not a universal endpoint worst case.

A.12.3 A CORRELATED EXTENSION OF THE LOCAL MINIMAX BUDGET

Retain the aligned second moments of Eq. (A.69) with fixed $-1 < \rho < 1$, but allow non-Gaussian joint endpoint laws. Independence of endpoints is now replaced by that cross-covariance constraint; the source remains Gaussian. Let $\theta = \arccos \rho$. The same projection argument bounds the normalized local deficit by the jointly Gaussian value, whose scalar integral is

$$I_\rho(r) := \int_0^1 \frac{r(1-\rho^2)}{[(1-t)^2 r + t^2 + 2\rho\sqrt{r}\, t(1-t)]^2}\, dt$$
$$= \left(\frac{\theta}{\sin\theta} - \rho\right) \cosh\left(\frac{1}{2}\log r\right) + 1 - \rho\frac{\theta}{\sin\theta}. \tag{A.116}$$

To verify the integral, use the polar angle ϕ of $P_r(t)$ from Theorem A.5. Its derivative is $\sqrt{r}\sin\theta / D(t; r, \rho)$, and inversion gives

$$t(\phi) = \frac{\sqrt{r}\sin\phi}{\sin(\theta-\phi) + \sqrt{r}\sin\phi}, \qquad I_\rho(r) = \frac{1}{\sqrt{r}\sin\theta} \int_0^\theta \big[\sin(\theta-\phi) + \sqrt{r}\sin\phi\big]^2 d\phi.$$

The squared-sine integral is $(\theta - \sin\theta\cos\theta)/2$, and the mixed-sine integral is $(\sin\theta - \theta\cos\theta)/2$, which yield Eq. (A.116). Since $\theta/\sin\theta - \rho > 0$, $\sum_k I_\rho(r_k)$ is uniquely minimized by $A = B$. A jointly Gaussian law with an independent label attains this envelope. At $\rho \uparrow 1$ the penalty and the deficit vanish, recovering the deterministic affine-coupling boundary. Fixing an arbitrary cross covariance instead of a common normalized modal correlation would be a different optimization problem.

A.13 GUIDANCE (PROPOSITION 4)

A.13.1 THE INITIAL TRANSLATION

Derivation of the initial translation. For $t < 1$, the interpolation identity gives $v = (\widehat{x}_1 - x)/(1-t)$, where $\widehat{x}_1$ is the relevant posterior mean. Therefore $v_{\mathrm{c}} - v_{\mathrm{u}} = \big[\mathbb{E}[x_1 \mid x_t, y] - \mathbb{E}[x_1 \mid x_t]\big]/(1-t)$, and the tower property gives $\mathbb{E}[x_1 \mid x_t] = \mathbb{E}[\mathbb{E}[x_1 \mid x_t, Y] \mid x_t]$. At $t = 0$, x_0 is independent of Y, so $\mathbb{E}[x_1 \mid x_0, y] = \mathbb{E}[x_1 \mid y]$ and $\mathbb{E}[x_1 \mid x_0] = \mu$; the bracket is $\mu_y - \mu$, and $v_{\mathrm{u}}(x, 0) = \mu - x$. Hence $v_w(x, 0, y) = \mu - x + w(\mu_y - \mu)$. □

When $p(y' \mid x_t) \to \delta_y$ the bracket tends to zero and the guidance term becomes $\sum_{y' \neq y} p(y' \mid x_t)\big[\hat{x}_1(\cdot, y) - \hat{x}_1(\cdot, y')\big]$, a correction weighted by the conditions the state has not yet excluded.

A.13.2 THE CONTINUOUS-TIME EARLY-GUIDANCE COVARIANCE EXPANSION

Proof of Proposition 4. For fixed x and y, the Gaussian-source likelihood of $x_1 = z$ given $x_t = x$ is proportional to

$$(1-t)^{-d} \exp\Big[-\frac{1}{2}\Big(\frac{x - tz}{1-t}\Big)^\top A^{-1}\Big(\frac{x-tz}{1-t}\Big)\Big], \tag{A.117}$$

whose log has t-derivative $d - x^\top A^{-1}x + x^\top A^{-1}z$ at $t = 0$. Differentiating the normalised posterior expectation therefore gives

$$\partial_t\, \mathbb{E}[x_1 \mid x_t = x,\, Y = y]\big|_{t=0} = B_y A^{-1} x, \tag{A.118}$$

and $BA^{-1}x$ for the unconditional posterior. Since $v_{\mathrm{c}} = (\hat{x}_1 - x)/(1-t)$, it follows that at $t = 0$

$$v_w(x, 0, y) = \mu + w(\mu_y - \mu) - x, \qquad D_x v_w(x, 0, y) = -I, \tag{A.119}$$
$$\partial_t v_w(x, t, y)\big|_{t=0} = \mu + w(\mu_y - \mu) - x + C_y A^{-1} x, \qquad C_y = (1-w)B + wB_y. \tag{A.120}$$

The material derivative along the flow is therefore $\ddot{X}^{(w)}(0) = C_Y A^{-1} x_0$, the translation terms cancelling, and the L^2 Taylor expansion is

$$X^{(w)}(t) = (1-t)x_0 + t\big[\mu + w(\mu_Y - \mu)\big] + \frac{t^2}{2} C_Y A^{-1} x_0 + o_{L^2}(t^2). \tag{A.121}$$

Independence of x_0 and Y with $\mathbb{E}x_0 = 0$ removes the cross terms between x_0 and the condition mean, and the symmetry of C_Y gives

$$\mathrm{Cov}\left(X^{(w)}(t)\right) = (1-t)^2 A + t^2\,\mathbb{E}C_Y + t^2 w^2 V + o(t^2), \tag{A.122}$$

$$\mathbb{E}C_Y = (1-w)B + wW = B - wV. \tag{A.123}$$

Substituting gives $\mathrm{Cov}(X^{(w)}(t)) = (1-t)^2 A + t^2[B + w(w-1)V] + o(t^2)$, i.e. Eq. (10); projecting on u gives the level form. The regularity assumed (differentiation under the conditional expectations and a second-order L^2 expansion of the ODE at 0) holds, for example, for finite Gaussian conditional families with positive-definite covariances and finite parameters. □

Remark A.7 (boundaries). For $w > 1$ and any direction with positive between-condition variance, Proposition 4 gives a positive leading covariance excess over the prescribed path even under exact integration. It neither says that an early state already exceeds the endpoint data variance nor proves that the excess survives to $t = 1$; the measured window length, endpoint persistence, trained-model behaviour and the best w_0 still require the empirical evidence of Section 4.

A.13.3 The first Euler step as a discrete counterpart

Corollary A.3 (first-step level variance)**.** *Let $\ell_0 = u^\top x_0$, $a_{\mathrm{lev}} = u^\top A u$ and $V_{\mathrm{between}} = \mathrm{Var}_Y(u^\top \mu_Y) = u^\top V u$. One Euler step of length h on the initial guided field derived in Appendix A.13.1 gives $\ell_h = (1-h)\ell_0 + h\,u^\top \mu + hw\,u^\top(\mu_Y - \mu)$ and*

$$\mathrm{Var}_{x_0,Y}(\ell_h) = (1-h)^2 a_{\mathrm{lev}} + h^2 w^2 V_{\mathrm{between}}. \tag{A.124}$$

Proof. This is one Euler step on $\mu - x + w(\mu_y - \mu)$ projected on u; $\mathrm{Cov}(\ell_0, u^\top \mu_Y) = 0$ because x_0 and Y are independent. □

The $h^2 w^2 V$ term is the mean-translation contribution alone; the continuous-time excess over the reference path, which also contains the conditional covariance dynamics, has coefficient $w(w-1)$, and the two must not be conflated. On HumanML3D the between-condition variance is at least a quarter of the total and sits mostly in the level, so with $w = 5$–10 neither term is small.

A.13.4 How fast the state becomes informative

For the homoscedastic Gaussian family $x_1 \mid y \sim \mathcal{N}(\mu_y, W)$, let $V = \mathrm{Cov}(\mu_Y)$ and $\Sigma_{t|y} = (1-t)^2 A + t^2 W$. The expected log-likelihood ratio for a fixed pair (y, y'), with x_t drawn conditionally on y, is the Gaussian KL divergence $\frac{t^2}{2}\|\mu_y - \mu_{y'}\|^2_{\Sigma^{-1}_{t|y}}$. Averaging over independent Y, Y' gives $t^2\,\mathrm{tr}(\Sigma^{-1}_{t|y} V)$, with leading term $t^2\,\mathrm{tr}(A^{-1}V)$ as $t \to 0$. If A and V share an eigenbasis, the trace is $\sum_k v_k/a_k$, so information can accumulate across directions. The HumanML3D bandwise variance measurements motivate this mechanism, but do not establish a matrix inequality $W \preceq 0.75B$ or a quantitative posterior-concentration time. The initial window is measured on the learned field in Figure 4a.

A.13.5 A one-dimensional Gaussian diagnostic

Consider the level coordinate alone: $x_1 \mid y \sim \mathcal{N}(z_y, b^*)$, $z_y \sim \mathcal{N}(0, \tau^2)$, marginal variance $b_m = b^* + \tau^2$, and source $\mathcal{N}(0, a)$; both fields are linear and the guided flow is affine in (x_0, z_y). Solving the guided ODE exactly, constant $w = 7.5$ gives an endpoint level βz_y with $\beta = 2.4$ when $b^* = \tau^2$ (and within-condition variance $b_m^{1-w} b^{*w}$), $\beta = 1.14$ when $b^* = 0.1\,b_m$, and $\beta = 1.01$ when $b^* = 0.01\,b_m$; lowering w for $t < 0.1$ changes β by at most 15%, and 24-step Euler agrees with the exact solution. This scalar diagnostic does not reproduce the measured sensitivity to early guidance in Figure 4. Information from other coordinates is a possible explanation, as suggested by the trace calculation above; these scalar calculations do not establish that high dimensionality is necessary.

A.14 Reproducible checks on exact population fields

These checks evaluate the stated formulas and numerical integrators directly; they do not use training data, neural networks, or model-quality metrics. We use 65-decimal-digit arithmetic for scalar Runge–Kutta responses, symbolic algebra for the Heun coefficients, and double precision for the matrix checks. The proofs above establish the results; these computations check coefficients, convergence orders, and non-commuting implementations.

Table 6: **Endpoint cancellation at the matched covariance.** Absolute normalized endpoint error $|R_{\mathcal{M},N}(1) - 1|$ for the exact Gaussian field under independent coupling and a linear path. The last column is $\log_2(e_{64}/e_{128})$.

method	$N = 32$	$N = 64$	$N = 128$	observed order
Heun	2.0419×10^{-5}	2.5536×10^{-6}	3.1924×10^{-7}	2.9998
RK4	2.6219×10^{-9}	8.1987×10^{-11}	2.5625×10^{-12}	4.9998

To check the extension beyond a linear path and uniform grid, set $\rho = 0.4$, $s(t) = (1 - \cos \pi t)/2$, $c(t) = 1 + 0.3 \sin^2 \pi t$, $\alpha(t) = c(t)(1 - s(t))$, and $\beta(t) = c(t)s(t)$, with grid $t_n = (1 - \cos(n\pi/N))/2$. At $r = 1$, the observed endpoint orders from $N = 64$ to 128 are 2.9962 for Heun and 4.9955 for RK4, consistent with Theorem A.4. Direct root finding gives $r_{24} = 1.1183094508442$ for independent-coupling Heun.

Symbolic checks recover $g' + g^2 = r(1-\rho^2)/D^2$, the Heun identity $a_3 = k'/8$, the Euler–Maclaurin correction in Eq. (A.79), and its matched coefficient $-3(4+\pi)/32$. With fixed random seed 1907, 40 independent pairs of four-dimensional positive-definite covariances at $N = 2, 4, 24$ verify the direct Euler covariance against the modal formula and place actual Gaussian W_2 between the trace floor and the mode-pairing upper bound. Thirty heterogeneous, generally non-commuting conditional families verify the budget decomposition and the optimizer of Proposition A.5; further direct updates use a singular target covariance. Scalar quadrature checks Eq. (A.116), and the correlated equal-angle grids attain Eq. (A.81). Appendix A.11.7 reports the extended-precision search for κ_N and checks the conditional sandwich on empirical HumanML3D families. The searched constants are numerical estimates rather than certified suprema; the global bound and its order are supplied by Lemma A.4.

A.15 Numerical robustness under covariance shift

We quantify the sensitivity of source design to changes in target covariance and show that matching a reference covariance is the unique minimax fixed source over a multiplicative covariance uncertainty set. Fix a length supported by the source table and one channel of dimension d. Write $B_0 \succ 0$ for the training covariance and $A \succ 0$ for the source used consistently in training and sampling. The endpoint results assume independent coupling, linear interpolation, $N \geq 2$ uniform Euler steps, and the exact conditional population field with $w = 1$; each source has its own corresponding field. Learned-field error is treated separately below.

A.15.1 Source support and second-order covariance-shift cost

A full-rank Gaussian source has positive density throughout $\mathbb{R}^d$, so covariance calibration imposes no hard support restriction on target laws. More precisely, if an unrestricted conditional transport map $F_I(\cdot, y)$ pushes $\mathcal{N}(0, I)$ to a target P_y, then $F_A(x, y) = F_I(A^{-1/2}x, y)$ pushes $\mathcal{N}(0, A)$ to the same target; the reverse construction composes with $A^{1/2}$. Thus the two sources have the same expressible target laws for unrestricted maps. This is an expressivity statement, not equivalence of training with a fixed neural architecture. TSCM also keeps $p(x_0 \mid y, L) = p(x_0 \mid L)$, so the source is shared across conditions at a fixed length.

The block-diagonal Gaussian $q_A = \mathcal{N}(0, \mathrm{diag}(A_c))$ is the maximum-entropy density with zero mean and the prescribed within-channel covariances. Indeed, for any density q satisfying those constraints and having finite entropy, $h(q) = h(q_A) - D_{\mathrm{KL}}(q\|q_A) \leq h(q_A)$, because the expectation of the block-diagonal quadratic form in $\log q_A$ is fixed. The source therefore specifies temporal second moments without additionally specifying higher-order structure or cross-channel dependencies.

For a condition y, let its Gaussian target be $\mathcal{N}(\mu_y, C_y)$, with $C_y \succ 0$, and define $E_N(A; y) = W_2(P^y_{N,A}, \mathcal{N}(\mu_y, C_y))$, where $P^y_{N,A}$ is the Euler endpoint law. Retain $\varepsilon_N = 1 - R_N(1)$ and c_N from Corollary A.1.

Corollary A.4 (combined estimation and covariance-shift error)**.** *Suppose*

$$e^{-\eta} B_0 \preceq A \preceq e^{\eta} B_0, \qquad e^{-\gamma_y} B_0 \preceq C_y \preceq e^{\gamma_y} B_0, \qquad \eta, \gamma_y \geq 0.$$

With $\delta_y = \eta + \gamma_y$*,*

$$\varepsilon_N \sqrt{\operatorname{tr} C_y} \leq E_N(A; y) \leq [1 - R_N(e^{\delta_y})] \sqrt{\operatorname{tr} C_y}. \tag{A.125}$$

Consequently,

$$0 \leq \frac{E_N(A; y) - \inf_{S \succ 0} E_N(S; y)}{\sqrt{\operatorname{tr} C_y}} \leq R_N(1) - R_N(e^{\delta_y}) = c_N \delta_y^2 + O_N(\delta_y^4). \tag{A.126}$$

No commutation assumption or bound on $\|\mu_y\|$ *is required.*

Proof. The affine population field is $v^\star_A(x, t, y) = \mu_y + G_{A,C_y}(t)(x - t\mu_y)$, with $G_{A,C_y}(t) = [tC_y - (1-t)A][(1-t)^2 A + t^2 C_y]^{-1}$. For $Z_n = X_n - t_n \mu_y$, Euler gives $Z_{n+1} = [I + N^{-1} G_{A,C_y}(t_n)] Z_n$, hence $\mathbb{E} X_n = t_n \mu_y$ exactly. The two covariance comparisons imply $e^{-\delta_y} C_y \preceq A \preceq e^{\delta_y} C_y$. Thus every eigenvalue r_k of $C_y^{-1/2} A C_y^{-1/2}$ lies between $e^{-\delta_y}$ and e^{δ_y}. The mode-pairing upper bound of Proposition 3, with positive weights summing to $\operatorname{tr} C_y$, and the monotonicity of Theorem A.1 give the upper bound in Eq. (A.125). Corollary A.2 gives its lower bound and attainment at $S = C_y$ for the actual endpoint W_2. The even expansion of Corollary A.1 then yields Eq. (A.126). □

The corresponding modal gain obeys $\Lambda(r_k) \leq \cosh(\delta_y/2) = 1 + \delta_y^2/8 + O(\delta_y^4)$. Both quantities respond only quadratically to a small log-spectral shift. For example, the exact $N = 24$ normalised error floor is 0.05218; ratios in $[1/2, 2]$ or $[1/4, 4]$ give upper bounds 0.05407 or 0.05993. These are Gaussian population-field endpoint bounds, not FID predictions. The absence of μ_y from the bounds separates mean extrapolation by the learned model from covariance-induced numerical error. A worst-mode spectral bound may be conservative near low-variance modes; the variance-weighted risk in Appendix B.7 remains the relevant complementary diagnostic.

A.15.2 The minimax fixed source under unknown covariance

Define the multiplicative uncertainty set and normalised endpoint loss by

$$\begin{aligned} \mathcal{U}_\gamma(B_0) &= \{B \succ 0 : e^{-\gamma} B_0 \preceq B \preceq e^{\gamma} B_0\}, \\ \ell_N(A, B) &= \frac{W_2(P_{N,A,B}, \mathcal{N}(0, B))}{\sqrt{\operatorname{tr} B}}, \end{aligned} \tag{A.127}$$

where $P_{N,A,B}$ is the Euler endpoint law for the population field associated with source $\mathcal{N}(0, A)$ and target $\mathcal{N}(0, B)$. The set permits covariances with different eigenvectors.

Theorem A.8 (unique minimax source under covariance uncertainty)**.** *For every* $B_0 \succ 0$*,* $\gamma \geq 0$*, and* $N \geq 2$*,*

$$\inf_{A \succ 0} \sup_{B \in \mathcal{U}_\gamma(B_0)} \ell_N(A, B) = 1 - R_N(e^{\gamma}), \qquad A^\star = B_0 \quad \textit{uniquely}. \tag{A.128}$$

This is a minimax statement for actual Gaussian endpoint W_2*, including non-commuting candidate sources and targets.*

Proof. For $A = B_0$, Corollary A.4 with $\eta = 0$ gives $\ell_N(B_0, B) \leq 1 - R_N(e^{\gamma})$ throughout the uncertainty set. At $B = e^{\gamma} B_0$, all generalised ratios equal $e^{-\gamma}$ and the output covariance is $R_N(e^{\gamma})^2 B$. The Gaussian distance therefore attains the bound.

For any $A \neq B_0$, diagonalise $B_0^{-1/2} A B_0^{-1/2} = U \operatorname{diag}(e^{s_i}) U^\top$. Set $\sigma_i = 1$ for $s_i \geq 0$ and $\sigma_i = -1$ otherwise, and choose

$$B_A = B_0^{1/2} U \operatorname{diag}(e^{-\gamma \sigma_i}) U^\top B_0^{1/2} \in \mathcal{U}_\gamma(B_0). \tag{A.129}$$

An invertible congruence simultaneously reduces A and B_A to $\mathrm{diag}(e^{s_i})$ and $\mathrm{diag}(e^{-\gamma\sigma_i})$. Their generalised ratios are thus $r_i = e^{s_i+\gamma\sigma_i}$, with $|\log r_i| = |s_i| + \gamma$. Writing $c_\gamma = R_N(e^\gamma)$, strict monotonicity gives $R_N(r_i) \le c_\gamma$ for every i, with a strict inequality for at least one i because $A \ne B_0$. The endpoint covariance formula $C_{\mathrm{out}} = B_A^{1/2} R_N(B_A^{-1/2} A B_A^{-1/2})^2 B_A^{1/2}$ therefore implies $C_{\mathrm{out}} \preceq c_\gamma^2 B_A$ and $\operatorname{tr} C_{\mathrm{out}} < c_\gamma^2 \operatorname{tr} B_A$. The trace lower bound of Eq. (A.73) yields

$$\ell_N(A, B_A) \ge 1 - \sqrt{\frac{\operatorname{tr} C_{\mathrm{out}}}{\operatorname{tr} B_A}} > 1 - c_\gamma.$$

Every other source has a strictly larger worst-case loss, proving uniqueness. □

Matching training covariance is therefore a robust fixed-source design when the test covariance is unknown within this relative uncertainty set. An estimated source with $e^{-\eta} B_0 \preceq \widehat{A} \preceq e^{\eta} B_0$ satisfies

$$\sup_{B \in \mathcal{U}_\gamma(B_0)} \ell_N(\widehat{A}, B) \le 1 - R_N(e^{\gamma+\eta}). \tag{A.130}$$

If $B_0 \ne I$, white noise is not minimax on this set. There is also a pointwise advantage over white noise in some neighbourhood of B_0: it is strict at $B = B_0$ by Corollary A.2, and persists by continuity of the finite-step Gaussian endpoint loss on positive-definite covariances. This local comparison does not assert dominance at every covariance in an arbitrarily large uncertainty set.

Corollary A.5 (non-Gaussian local robustness under covariance uncertainty)**.** *With $\mathcal{P}(B)$ and $\mathcal{J}$ as in Appendix A.12,*

$$\inf_{A \succ 0} \sup_{B \in \mathcal{U}_\gamma(B_0)} \sup_{P \in \mathcal{P}(B)} \mathcal{J}(A; P) = d + \frac{\pi d}{2} \cosh(\gamma/2), \qquad A^\star = B_0 \quad \textit{uniquely}. \tag{A.131}$$

Proof. Theorem A.7 identifies the inner supremum as $d + \frac{\pi}{2} \sum_i \cosh(\frac{1}{2} \log r_i)$. At $A = B_0$, all $|\log r_i| \le \gamma$, and equality is attained by $B = e^\gamma B_0$ with a Gaussian target and an independent condition. For $A \ne B_0$, the covariance B_A in Eq. (A.129) gives $|\log r_i| = |s_i| + \gamma$, with at least one strict increase. Strict increase of $\cosh$ on $[0, \infty)$ proves uniqueness. □

The worst local-budget increase is $\frac{\pi d}{2}[\cosh(\gamma/2) - 1] = \frac{\pi d}{16}\gamma^2 + O(\gamma^4)$. This result permits arbitrary changes in higher-order target structure and condition information within $\mathcal{P}(B)$. It controls the integrated normalised one-step covariance deficit, with each step starting from the true interpolation law; it is not a non-Gaussian endpoint theorem.

A.15.3 Test-condition averages and learned-field error

Per-condition spectral comparability is one way to express transferable temporal structure; it does not follow from similarity of marginal covariances. An alternative averages over the test-condition law q. For Gaussian conditional targets with finite $\mathbb{E}_q \|C_Y\|_F^2$ and $\mathbb{E}_q \|\mu_Y\|^2$, set $W_q = \mathbb{E}_q C_Y$, $\mathcal{H}_q = \mathbb{E}_q[(C_Y - W_q)^2]$, and $\mathcal{R}_{N,q}(A) = \mathbb{E}_q E_N(A; Y)^2$. The pointwise bounds in the proof of Theorem A.6 give, for every shared $A \succ 0$,

$$\begin{aligned} 0 &\le \mathcal{R}_{N,q}(A) - \inf_{S \succ 0} \mathcal{R}_{N,q}(S) \\ &\le \kappa_N \left[\operatorname{tr}\left((A - W_q) A^{-1} (A - W_q) \right) + \operatorname{tr}(A^{-1} \mathcal{H}_q) \right]. \end{aligned} \tag{A.132}$$

Indeed, expand $A - C_Y = (A - W_q) - (C_Y - W_q)$ in the conditional budget; the cross terms vanish on averaging. Subtracting the common lower floor $\varepsilon_N^2 \operatorname{tr} W_q$ proves the claim. This bound permits singular conditional covariances and requires neither commutation nor a uniform spectral lower bound. It separates displacement of the test within-condition average covariance from covariance heterogeneity. Changing only the conditional means, while retaining the covariance family and its weights, does not change this population-field numerical risk.

To connect these guarantees to a trained model, let $\widehat{P}^y_{N,A}$ be its Euler endpoint law and couple the learned and population Euler recursions through the same source draw. Write $h = 1/N$, $t_n = nh$, and

$$a_n = \left(\mathbb{E} \| v_\theta(\widehat{X}_n, t_n, y) - v_A^\star(\widehat{X}_n, t_n, y) \|^2 \right)^{1/2}.$$

If $v_A^\star(\cdot, t_n, y)$ is L_n-Lipschitz and $a_n < \infty$, then the triangle inequality and the coupled Euler recursion give

$$\begin{aligned} W_2(\widehat{P}^y_{N,A}, p_\star(\cdot \mid y)) &\le W_2(P^y_{N,A}, p_\star(\cdot \mid y)) + W_2(\widehat{P}^y_{N,A}, P^y_{N,A}) \\ &\le W_2(P^y_{N,A}, p_\star(\cdot \mid y)) + \sum_{n=0}^{N-1} h a_n \prod_{j=n+1}^{N-1} (1 + hL_j). \end{aligned} \tag{A.133}$$

For completeness, if $e_n = (\mathbb{E}\|\widehat{X}_n - X_n\|^2)^{1/2}$, Minkowski's inequality gives $e_{n+1} \le (1+hL_n)e_n + ha_n$ and $e_0 = 0$; iteration bounds the Wasserstein distance by the stated sum. For the Gaussian conditional target of Corollary A.4, the first term is at most $[1 - R_N(e^{\eta+\gamma_y})]\sqrt{\operatorname{tr} C_y}$. Thus covariance transfer bounds the numerical term; accuracy of the trained conditional field enters separately through the errors a_n. More explicitly, defining $D_N(A; y) = W_2(\widehat{P}^y_{N,A}, P^y_{N,A})$, the condition $E_N(I; y) - E_N(A; y) > D_N(I; y) + D_N(A; y)$ is sufficient for the trained TSCM model to have smaller endpoint W_2 than the model trained with white noise, by the triangle and reverse triangle inequalities.

These results quantify the numerical consequences of covariance shift under the stated population-field assumptions. The endpoint statements apply channel-wise to the block-diagonal Gaussian surrogate, and the local budgets add across channels. They do not claim optimality over arbitrary cross-channel source covariances. Solver-dependent scales, additional CFG bias, and unsupported lengths retain the separate roles described in Appendices A.8 and A.13 and Section 3.1. Appendix B.7 tests transfer to a held-out marginal reference. Appendix A.11.8 resolves the held-out data into caption clusters and finds lower covariance-only oracle risk with the training-derived TSCM source than with white noise in all 56 cluster surrogates. These diagnostics provide empirical support for the numerical transfer argument; they do not evaluate semantic OOD generation.

Table 7: **Additional within-model text-to-motion results.** The diffusion and latent white/TSCM pairs isolate source replacement at constant guidance. The DiT-L rows include an early-guidance intervention, so the white-to-TSCM difference combines source and schedule changes. $\pm$ is the 95% confidence interval over 20 test generations.

			R-precision↑			FID↓	Diversity↑
dataset	model	source / schedule	top-1	top-2	top-3		
HumanML3D	DiT-M x_0 diffusion	white	$0.534^{\pm.002}$	$0.727^{\pm.002}$	$0.821^{\pm.002}$	$0.173^{\pm.003}$	$9.953^{\pm.117}$
		TSCM	$0.521^{\pm.002}$	$0.721^{\pm.002}$	$0.819^{\pm.002}$	$0.108^{\pm.003}$	$9.757^{\pm.109}$
	DiT-M, T/4 latent	white	$0.545^{\pm.002}$	$0.741^{\pm.002}$	$0.833^{\pm.001}$	$0.256^{\pm.004}$	$9.944^{\pm.094}$
		TSCM	$0.543^{\pm.002}$	$0.730^{\pm.002}$	$0.822^{\pm.001}$	$0.097^{\pm.003}$	$9.823^{\pm.101}$
		TSCM + early guidance	$0.541^{\pm.003}$	$0.730^{\pm.002}$	$0.823^{\pm.002}$	$0.083^{\pm.002}$	$9.819^{\pm.096}$
	DiT-L (CLIP only)	white	$0.529^{\pm.002}$	$0.721^{\pm.002}$	$0.818^{\pm.002}$	$0.386^{\pm.003}$	$10.035^{\pm.109}$
		TSCM-Toeplitz + early guidance	$0.544^{\pm.002}$	$0.732^{\pm.001}$	$0.824^{\pm.002}$	$0.052^{\pm.001}$	$9.617^{\pm.108}$
		TSCM + early guidance	$0.539^{\pm.002}$	$0.730^{\pm.002}$	$0.821^{\pm.001}$	$0.047^{\pm.001}$	$9.616^{\pm.104}$

B Experimental details

B.1 Training recipe

Data and metrics. HumanML3D (Guo et al., 2022): 24,546 training clips (with mirrored copies) of 40–196 frames at 20 fps in the standard 263-dimensional per-frame representation, standardised per channel; the official validation and test splits and the evaluation protocol of Guo et al. (2022) (FID, R-precision, diversity).

Hardware. Unless stated otherwise, experiments run on NVIDIA GeForce RTX 5090 GPUs (32 GB) with PyTorch 2.11 and CUDA 12.8, using one GPU per model, bf16 autocast, and float32 source draws. The separate sampling-cost measurements in Appendix B.5 use an A100.

Model and training. DiT-M: 12 blocks, hidden 512, 8 heads, MLP 2048, one token per frame with a learned positional encoding (dropout 0.1), timestep embedding and CLIP ViT-B/32 pooled text embedding added and injected by adaptive layer norm; zero-initialised modulation and output

layers; 10% text-condition dropout during training. AdamW ($\beta = (0.9, 0.99)$, weight decay 10^{-5}), learning rate 2×10^{-4}, 2000 warm-up updates, $\times 0.1$ at update 50,000, batch 64, bf16 autocast, EMA 0.9999 (the EMA weights are evaluated), 400 epochs ($\approx$ 153k updates), training seeds 3407, 3408 and 3409 for the controlled source study. Sequences are cropped to their length and padded to 196 frames with a mask. Sampling: 24 uniform Euler steps in float32 for the source draw. DiT-L backbone: 16 blocks, hidden 768, 12 heads, MLP 3072 (172M parameters), FFN and attention dropout 0.1, condition dropout 0.2, weight decay 0.01 on linear weights, cosine decay to zero. The CLIP-only variant in Table 7 trains for 800 epochs. The DiT-L rows in the main Table 3 use the same backbone with a T5-large (Raffel et al., 2020) sentence vector (masked mean over up to 64 tokens, 1024-d, projected and added to the same adaptive-layer-norm condition) beside the CLIP embedding, with per-example modality dropout (both 0.4, T5 only 0.2, CLIP only 0.2, none 0.2) and 1000 scheduled epochs. In our runs the CLIP-only DiT-L overfits the captions; the second encoder and modality dropout present each caption in four conditioning views and mitigate this overfitting. White and TSCM sources within the main-table DiT-L pair use this same conditioning and training recipe. KIT-ML (Plappert et al., 2016) (251-dimensional representation, 4,394 training clips of 24–196 frames at 12.5 fps, official splits and evaluator).

B.2 Evaluation protocol

For the DiT-M flow experiments, every epoch the model generates the validation split once at its configured guidance scale (the swap models of Table 8 also at $w = 5.5$), and the best checkpoint per scale is kept together with the final EMA checkpoint. For the sixteen-source controlled study, all sources and all three training seeds use the same validation epochs (201–400), checkpoint-search budget, 24-step Euler sampler, and FID evaluator. Checkpoint selection uses validation FID among the retained candidates, with differences below 0.01 retaining the earlier checkpoint. We freeze the selected checkpoint for each seed before scanning constant guidance. Every source then uses the same discrete descent procedure: start from the same scale, compare neighboring scales with the same step rule, and stop when neither neighbor lowers validation FID. At each tested scale, each seed is evaluated separately with one validation generation from its fixed checkpoint; the three validation FIDs are averaged, and the best scale found by this search is used for that source. Figure 7 plots the tested scales and three-seed mean FIDs at the selected checkpoints; stars mark the scales used for Table 8. The schedule grid ($T_0 \in \{1, 2, 3, 4\}$ steps $\times$ $w_0 \in \{2, 3\}$ $\times$ $w_1 \in \{$best, one notch above$\}$) is then swept the same way on the selected checkpoint. Each trained model is evaluated using 20 independent generations of the test split at the selected setting. For the three-seed controlled source study (Tables 8 and 9), each run is evaluated with its validation-selected checkpoint and the constant source-specific guidance scale. We first average the 20 sampling repeats within each seed, then report the mean and sample standard deviation of those three per-seed means. Validation selection and evaluation use the official evaluator checkpoint of Guo et al. (2022). The step counts and guidance schedules for the other generators are given in their respective subsections; the grid above is specific to DiT-M. Sampling confidence intervals in Tables 3 and 7 describe repeated generation from a fixed checkpoint, not training-seed variation. The main-table HumanML3D DiT-M white/TSCM results ($0.471/0.116$) use separate single training runs with their own validation selection; they are not the three-seed aggregates ($0.497/0.108$) of Table 8.

B.3 The sources

We define the following sources per channel c and length L, sampled through their Cholesky factors. Stationary kernel and ACF matrices use the leading $L \times L$ block of their 196×196 construction; length-dependent level modifications and TSCM are defined below: *white noise* I; the three *TSFlow kernel sources* $K' = K + K_{\text{white}}$, with $K_{\text{white}} = I$ and $K_{ij} = e^{-\tau}$ (OU), $e^{-\tau^2}$ (squared-exponential), or $e^{-\sin^2 \tau}$ (periodic), where $\tau = \pi|i-j|/20$; the *uncentered Toeplitz* source $A_{\text{U}} = \text{Toep}(\rho_c^{\text{F}})$ with ρ^{F} the Bartlett autocorrelation of the uncentered standardised training sequences; the *centered Toeplitz* source $A_{\text{C}} = \text{Toep}(\rho_c^{\text{T}})$ with ρ^{T} from sequence-centred training sequences (both normalised to lag-0 $= 1$); the *level-rescaled* sources $HA_{\text{C}}H + \gamma\, d^{\text{T}} uu^\top$ with $\gamma = \frac{1}{4}$ and 4, the latter with and without a trace normalisation $\times L/(L + (\gamma - 1)d^{\text{T}})$ (the $\gamma = \frac{1}{4}$ source is trace-normalised too); the *time-permuted* source, A_{C} with a fixed random permutation of the time index per length; the four corners of the *level–fluctuation swap* with zero cross block, (T,T) $HA_{\text{C}}H + d^{\text{T}} uu^\top$, (F,T) $HA_{\text{U}}H + d^{\text{T}} uu^\top$, (F,F) $HA_{\text{U}}H + d^{\text{F}} uu^\top$ and (T,F) $HA_{\text{C}}H + d^{\text{F}} uu^\top$

= TSCM-Toeplitz, where $d^{\mathrm{T/F}} = u^\top A_{\mathrm{C/U}} u$ at the sequence's length; the *half-raised level* source $A_{\mathrm{C}} + \frac{1}{2}[\hat{d}_L - d^{\mathrm{T}}]_+ uu^\top$ with $\hat{d}_L$ the direct $L\,\mathrm{Var}(\mathrm{mean})$ table; and TSCM, the length-indexed shrunk second moment of Eq. (8), tabulated at $L = 40, 44, \ldots, 196$ and sampled at the sequence's own length (the leading block of the next tabulated length otherwise). All statistics come from the 24,546 training clips; the Bartlett estimator makes every Toeplitz block positive definite without jitter.

Table 8: **The sixteen Gaussian sources of the controlled study, same DiT-M, same recipe.** HumanML3D test FID, mean over three training seeds ± the standard deviation across seeds (each seed: 20 sampling repeats; per-seed values in Table 9), at 24 Euler steps and the source-specific, validation-selected constant guidance scale w listed below (sweeps in Figure 7). S_{24} is the oracle risk of Eq. (5). Level = $u^\top A_c u$ at 64 frames (data: 37); T/F denotes the centered/uncentered estimate of a component, A_{U}/A_{C} the Toeplitz matrices of the uncentered/centered autocorrelation, d^{F}/d^{T} their level powers.

source (per-channel covariance A_c)	level $u^\top A_c u$	S_{24}	w	FID↓
white noise, I	1	1.21	10	0.497 ± 0.070
TSFlow (OU, $e^{-\lvert\tau\rvert}$)	12.5	0.58	7.5	0.207 ± 0.004
TSFlow (SE, $e^{-\tau^2}$)	11.7	0.78	10	0.485 ± 0.026
TSFlow (periodic, $e^{-\sin^2\tau}$)	42	5.11	10	1.867 ± 0.284
centered Toeplitz, time-permuted	14	1.89	7.5	0.554 ± 0.018
uncentered Toeplitz, $\mathrm{Toep}(\rho^{\mathrm{F}})$	35	0.50	5.5	0.135 ± 0.002
centered Toeplitz, $\mathrm{Toep}(\rho^{\mathrm{T}})$	14	0.55	7.5	0.234 ± 0.010
centered Toeplitz, level $\times\frac{1}{4}$	4	0.69	10	0.431 ± 0.077
centered Toeplitz, level $\times 4$, trace-normalised	34	0.50	7.5	0.173 ± 0.013
centered Toeplitz, level $\times 4$	56	0.49	5.5	0.167 ± 0.003
centered Toeplitz, level raised halfway to the data's $\hat{d}_L$	28	0.48	5.5	0.141 ± 0.005
(T,T): $HA_{\mathrm{C}}H + d^{\mathrm{T}}uu^\top$	14	0.54	7.5	0.222 ± 0.015
(F,T): $HA_{\mathrm{U}}H + d^{\mathrm{T}}uu^\top$	14	0.54	7.5	0.218 ± 0.009
(F,F): $HA_{\mathrm{U}}H + d^{\mathrm{F}}uu^\top$	35	0.50	5.5	0.139 ± 0.008
(T,F) = TSCM-Toeplitz: $HA_{\mathrm{C}}H + d^{\mathrm{F}}uu^\top$	35	0.50	5.5	0.151 ± 0.017
TSCM: length-indexed second moment, Eq. (8)	36	0.46	5.5	$\mathbf{0.108 \pm 0.007}$

B.4 Training-seed variation of the controlled study

Every source of Table 8 was trained three times (seeds 3407, 3408, 3409; same recipe, 400 epochs, validation from epoch 201). Each run uses the checkpoint selected by the common validation-FID rule of Appendix B.2, with the final EMA checkpoint eligible in the same way for every source. The selected checkpoint is held fixed while the source-specific constant guidance scale is chosen by the common discrete descent search. We then evaluate each run at that scale using 20 sampling repeats. Table 9 lists the runs. The seed-to-seed standard deviation is 1.6–7.8% of the mean for twelve sources and 11–18% for four (white noise, level $\times\frac{1}{4}$, periodic, and TSCM-Toeplitz). TSCM has the lowest FID on every seed.

B.5 Cost of the source

TSCM adds no parameters and no learned module; its cost is a one-time statistic and a per-sample matrix product. The offline statistics (263 channels, the 40 crop lengths $L = 40, 44, \ldots, 196$, Ledoit–Wolf shrinkage, Cholesky factors) take 3,574 s of CPU time once with NumPy. At sampling time the source draw is a per-channel triangular product, costing at most $263L^2$ multiply–adds per sample against about 1.4×10^{11} for 24 guided DiT-M steps. Measured on one A100 at batch 32: the source draw takes 3.5 ms for a validation batch of mixed lengths (the implementation loops over the distinct lengths) and 0.7 ms when every length is 196; one guided-field evaluation (conditional and unconditional passes) takes 88 ms; 24 guided Euler steps take 2.12 s end to end, against 2.11 s from white noise on the same card. Across the mixed-length and fixed-length batches, the increase in median end-to-end time is 0.2–0.3% (computed before rounding), within the repeat-to-repeat

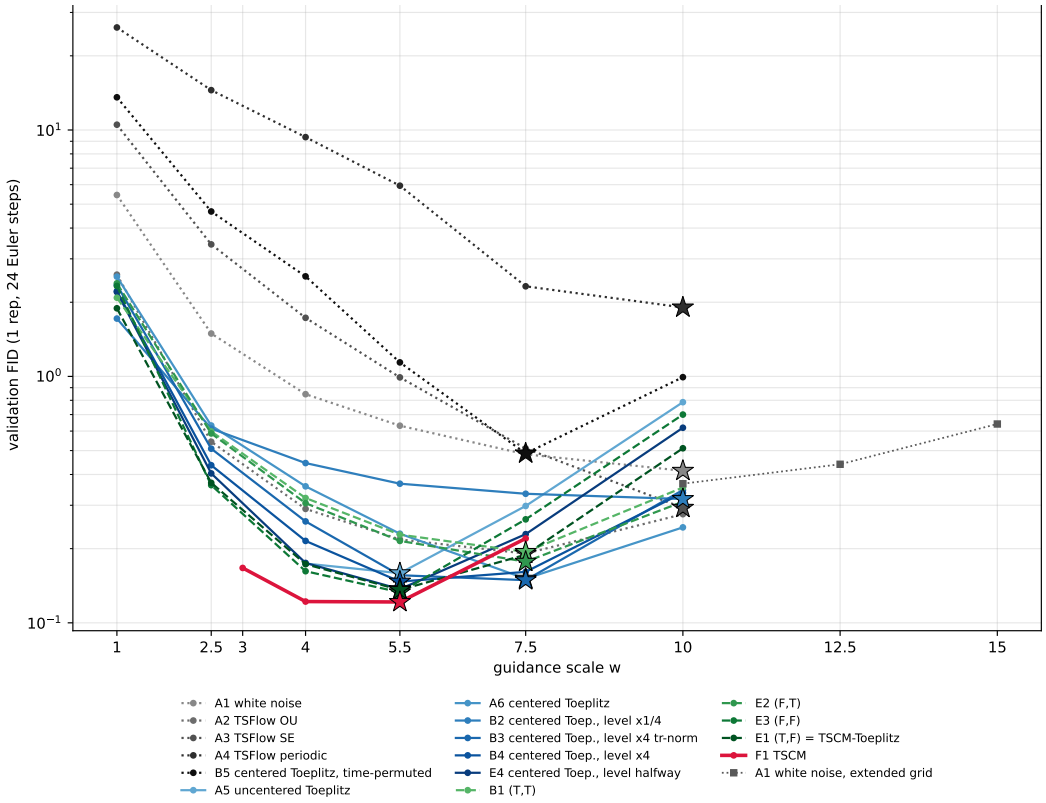

Figure 7: **Constant-guidance validation searches for all 16 sources.** HumanML3D validation FID at the guidance scales visited by the same discrete descent search, averaged over three training seeds (one validation generation per seed and visited w), using 24 Euler steps and the official evaluator. Each seed holds fixed its selected checkpoint during the shown search; stars indicate the best visited, source-specific constant scales used in Table 8. The table reports independent test generations.

Table 9: **Three training seeds per source.** HumanML3D test FID (mean over 20 sampling repeats) of each run and the mean ± sample SD over seeds, computed before rounding; S_{24} and w as in Table 8.

source	S_{24}	w	seed 3407	seed 3408	seed 3409	mean ± SD
white noise, I	1.21	10	.522	.418	.551	$.497 \pm .070$
TSFlow (OU)	0.58	7.5	.208	.211	.202	$.207 \pm .004$
TSFlow (SE)	0.78	10	.514	.478	.463	$.485 \pm .026$
TSFlow (periodic)	5.11	10	1.758	1.654	2.190	$1.867 \pm .284$
centered Toeplitz, time-permuted	1.89	7.5	.573	.552	.538	$.554 \pm .018$
uncentered Toeplitz	0.50	5.5	.137	.132	.135	$.135 \pm .002$
centered Toeplitz	0.55	7.5	.228	.230	.246	$.234 \pm .010$
centered Toeplitz, level $\times \frac{1}{4}$	0.69	10	.506	.435	.352	$.431 \pm .077$
centered Toeplitz, level $\times 4$, trace-normalised	0.50	7.5	.183	.179	.158	$.173 \pm .013$
centered Toeplitz, level $\times 4$	0.49	5.5	.170	.166	.163	$.167 \pm .003$
centered Toeplitz, level raised halfway	0.48	5.5	.137	.147	.141	$.141 \pm .005$
(T,T)	0.54	7.5	.231	.230	.204	$.222 \pm .015$
(F,T)	0.54	7.5	.227	.218	.208	$.218 \pm .009$
(F,F)	0.50	5.5	.130	.145	.142	$.139 \pm .008$
(T,F) = TSCM-Toeplitz	0.50	5.5	.134	.166	.154	$.151 \pm .017$
TSCM	0.46	5.5	.109	.115	.101	$\mathbf{.108 \pm .007}$

spread. The tabulated Cholesky factors occupy $682\,\mathrm{MB}$ of GPU memory as dense float32 ($676\,\mathrm{MB}$ on disk); packed triangular storage halves this, and loading one length at a time leaves at most about $40\,\mathrm{MB}$ resident.

B.6 The oracle risk and its correlation with FID

B_c is the temporal covariance of 12,000 random 64-frame training windows per channel, standardised, plus $10^{-3}\,\mathrm{tr}(B_c)/L$ on the diagonal as a noise floor. For every source, $r_{c,k}$ and $v_{c,k}$ are the eigenpairs of $B_c^{-1/2} A_c B_c^{-1/2}$ and S_{24} is Eq. (5). Over the 16 sources, S_{24} against log test FID gives Spearman 0.94 and log–log Pearson 0.93 (three-seed means; log–log Pearson 0.96 and Spearman 0.84 on the ten Toeplitz variants plus TSCM, whose seed-to-seed spread of 2–18% is comparable to their separation); sources that lack power where the data have it (white noise, the squared-exponential and periodic kernels, the permuted source, the level-reduced source) are predicted and measured to be bad, the Toeplitz family predicted and measured to be good, and TSCM is lowest on both axes; the cold part of S_{24} ($r_k < 1$) alone gives Spearman 0.92, the hot part gives Spearman -0.54; the minimum leave-one-source-out log–log Pearson correlation is 0.86. The oracle number of white-noise steps needed to match the centred-ACF source at 24 steps (exact Euler covariances, median channel) is 49–62 depending on the length, i.e. 2–2.6×. The level calibration oracle uses held-out windows (two folds by motion identity) and the training length histogram as weights: the centered Toeplitz source has length-weighted risk 0.0051; heating only its level to $\hat{d}_L$ gives 0.0037; matching all modes gives 0.0028; white noise gives 0.0157.

Distances to real validation windows. We additionally evaluate all 16 sources from Table 8 directly against real HumanML3D windows at $L = 64$. For each of the 263 channels, we compare 2,048 standardised validation windows in $\mathbb{R}^{64}$ with 2,048 source samples. Empirical W_2 is computed by exact linear assignment with squared Euclidean cost, followed by the square root of the mean squared distance across channels. The energy score uses the unbiased two-sample energy statistic, excluding within-sample self-pairs, followed by the square root of its channel average; unbiasedness applies to the underlying statistic, not the final square root. These scores compare temporal distributions within each channel. Real–real comparisons and $\mathcal{N}(0, \widehat{B}_{\mathrm{val}})$, a Gaussian matched to the estimated validation covariance, provide reference values (Table 10).

The periodic source is closer to real data than OU and white noise under both empirical distances, yet has substantially worse 24-step FID: greater proximity does not necessarily yield better few-step generation. Across the same 16 sources, S_{24} correlates more strongly with FID (Table 10b), including minimum leave-one-source-out log–log Pearson correlations of 0.86 versus 0.57 for empirical W_2. These empirical-distribution scores and the oracle's training-covariance score use different references.

Table 10: **Source proximity and prediction of trained-model FID.** (a) Empirical distances to real HumanML3D windows at $L = 64$, using 2,048 samples per side and aggregating over 263 channels. FID is the 24-step test FID averaged over three training seeds from Table 8. "Nearest" counts channels where a source has the smallest empirical W_2 among the 16 sources; reference rows are excluded. (b) Correlations with FID over the same 16 sources. Bold marks the best source distance or highest correlation.

(a) Source–data distances

source or reference	FID↓	empirical W_2 ↓	energy↓	nearest
real vs. real (validation–validation)	–	3.973	0.000	–
real vs. real (training–validation)	–	4.200	0.085	–
Gaussian $\mathcal{N}(0, \widehat{B}_{\mathrm{val}})$	–	6.212	0.792	–
TSCM	0.108	**6.173**	**0.775**	199
uncentered Toeplitz	0.135	6.904	1.020	21
(F,F)	0.139	6.926	1.022	0
level $\times 4$, trace-normalised	0.173	6.964	1.023	13
(F,T)	0.218	7.149	0.993	6
level raised halfway	0.141	7.187	1.066	9
TSCM-Toeplitz	0.151	7.548	1.129	0
centered Toeplitz	0.234	7.717	1.086	0
(T,T)	0.222	7.742	1.087	0
level $\times 4$	0.167	7.818	1.226	3
periodic	1.867	7.907	1.042	0
squared-exponential	0.485	8.297	1.118	10
OU	0.207	8.433	1.115	0
level $\times \frac{1}{4}$	0.431	8.561	1.138	2
time-permuted	0.554	9.433	1.125	0
white noise	0.497	10.786	1.184	0

(b) Association with 24-step FID

predictor	log–log Pearson	Spearman
empirical W_2	0.63	0.83
energy score	0.33	0.47
S_{24}	**0.93**	**0.94**

B.7 Estimation stability of the TSCM statistics

We re-estimate the TSCM statistics of Eq. (8) (same window law and shrinkage) from random subsets of the training clips drawn by motion identity (mirror pairs together) at budgets of 1%, 5%, 20% and 100% of the identities (8, 8, 4 and 1 draws), at $L = 64$ and $L = 196$, and score every estimate against one fixed reference covariance that no estimator sees: $\widehat{B}_{\mathrm{ref}}$ from 24,000 random 64-frame windows of the validation split, standardised, cross-sample mean removed, plus the noise floor of Appendix B.6. $\widehat{B}_{\mathrm{ref}}$ is itself an estimate, so the floor $[1 - R_{24}(1)]\sqrt{\operatorname{tr} \widehat{B}_{\mathrm{ref}}}$ is that of the reference, and the excess of the 100% estimate contains the reference's own noise. Table 11 reports the normalised risk $S_{24}/\sqrt{\operatorname{tr} \widehat{B}_{\mathrm{ref}}}$ summed over the 263 channels, its excess over the 100% estimate, the channel-median plain relative operator-norm error $\varepsilon = \|\widehat{A} - \widehat{A}_{100\%}\|_{\mathrm{op}} / \|\widehat{A}_{100\%}\|_{\mathrm{op}}$ against the 100% estimate at both lengths, the level power (channel medians) and the shrinkage intensity. This operator-norm error is plain, unwhitened; the whitened quantity is reported below. The same reference gives, at $L = 64$, floor 0.05218, TSCM 0.05235 ($+0.3\%$), TSCM-Toeplitz 0.05633 ($+7.9\%$), the uncentered and centered Toeplitz sources 0.05623 and 0.06284, the OU, squared-exponential and periodic kernels 0.0673, 0.0893 and 0.569, and white noise 0.1417 ($+172\%$); the TSCM, TSCM-Toeplitz, and white-noise excesses are quoted in Section 3.3.

Relative to the matched reference floor, the 1%, 5%, and 20% budgets give TSCM excesses of 2.4–2.9%, 0.9–1.2%, and 0.5%, respectively.

Table 11: **TSCM statistics estimated from subsets of the training clips, scored against a fixed validation-split reference ($L = 64$, $N = 24$).** Normalised risk over draws (mean $\pm$ SD) and its excess over the 100% estimate; channel-median plain, unwhitened relative operator-norm error $\varepsilon = \|\widehat{A} - \widehat{A}_{100\%}\|_{\mathrm{op}} / \|\widehat{A}_{100\%}\|_{\mathrm{op}}$ at 64 and 196 frames (range over draws; $\sqrt{L/n}$ in parentheses); level power; Ledoit–Wolf intensity. Floor 0.05218; TSCM-Toeplitz 0.05633; white noise 0.1417.

budget	windows 64 / 196	$S_{24}/\sqrt{\mathrm{tr}\,\widehat{B}_{\mathrm{ref}}}$	excess	plain ε, 64	plain ε, 196	level 64 / 196	ρ 64 / 196
1%	220–242 / 72–102	0.05357 ± 0.00013	0.0011–0.0014	0.19–0.34 (0.53)	0.37–0.48 (1.5)	30–36 / 59–78	0.04 / 0.11–0.14
5%	1.1k / 382–468	0.05274 ± 0.00006	0.0003–0.0005	0.09–0.14 (0.24)	0.17–0.23 (0.68)	31–37 / 74–92	0.01 / 0.03
20%	4.5k / 1.7k	0.05244 ± 0.00003	≤ 0.0001	0.04–0.05 (0.12)	0.08–0.09 (0.33)	35–38 / 94–102	0.003 / 0.008
100%	22.5k / 8.6k	0.05235	0	0 (0.05)	0 (0.15)	36.3 / 97.0	0.0005 / 0.0018

Three things follow. The risk converges to the floor with the budget: 1% of the identities (about 230 windows at $L = 64$) leaves an excess of 0.0011–0.0014 over the full estimate, 5% leaves 0.0003–0.0005, 20% ≤ 0.0001, and the full estimate is 0.00017 above the reference floor; even the 1% estimates (0.0534–0.0537) are below the full-data Toeplitz variant (0.0563), whose excess is structural and does not shrink with the budget. The plain relative operator-norm errors fall roughly like $\sqrt{L/n}$ and stay below it at every budget.

The whitened quantities of Corollary A.1 are not small for these estimates. For the following decomposition of the eight 1% draws, the reference is the full training estimate $F = \widehat{A}_{100\%}$ at the same length, rather than $\widehat{B}_{\mathrm{ref}}$; r_k and v_k are the eigenpairs of $F^{-1/2}\widehat{A}F^{-1/2}$ and $w_k = v_k^\top F v_k$. Against the 100% estimate, the spectral error $\delta = \max_k |\log r_k|$ has channel median 3.7–4.0 at $L = 64$ and 3.9–4.4 at $L = 196$, and the whitened operator norm $\max_k |r_k - 1|$ has channel median 39–52 and 47–78, so the bound of the corollary is vacuous at this budget. The modes responsible are the near-floor ones: at $L = 64$, 65–71% of the generalised modes have $|\log r_k| > 1$, but they carry 4–9% of the weight $\sum_k w_k$ and 6–11% of the risk sum, and their variance under the full estimate is 7×10^{-5} of the channel's largest eigenvalue (median); the normalised risk $S_{24}(\widehat{A}; F)/\sqrt{\mathrm{tr}\,F}$ itself is 0.0534 against the floor 0.0522. At $L = 196$, where the 1% draws have 72–102 windows, fewer than L, the same modes carry 41–48% of the weight and 49–61% of the risk, and the excess over the floor is 0.006–0.009.

The level power is the statistic that moves most with the budget (at 1%, 30–36 at 64 frames and 59–78 at 196, where the shrinkage toward μI pulls it down), consistent with Section 3.2: it is a long-window quantity and needs the most data. No model is retrained here; the study concerns the statistics only.

Residual means and covariance matching. The theoretical optimum matches $\mathrm{Cov}(x_1)$. For the affine Gaussian field under the linear path, the target mean contributes the trajectory $t\,m$, which Euler integrates exactly; it does not add $mm^\top$ to the optimal source covariance. HumanML3D motion inputs use training-set mean/standard-deviation normalisation. We measure the remaining mean on HumanML3D using the same one-window-per-eligible-clip law as Eq. (8), over all 40 supported lengths $L = 40, 44, \ldots, 196$. For each channel, let $\widehat{m}_{c,L} = n_L^{-1} \sum_i x_c^{(i)}$ and $\widehat{C}_{c,L} = \widehat{S}_{c,L} - \widehat{m}_{c,L}\widehat{m}_{c,L}^\top$, using the same windows and denominator n_L. This centres across windows at each temporal coordinate and preserves variation in window levels. The total mean-energy ratio is $e_L = \sum_c \|\widehat{m}_{c,L}\|^2 / \sum_c \mathrm{tr}\,\widehat{S}_{c,L}$; the per-channel level mean share is $q_{c,L} = (u^\top \widehat{m}_{c,L})^2 / (u^\top \widehat{S}_{c,L} u)$, and the level ratio is $\gamma_{c,L} = u^\top \widehat{S}_{c,L} u / (u^\top \widehat{C}_{c,L} u)$. Across the 40 lengths, e_L ranges from 0.24% to 1.77%, and the largest level ratio is below 1.09; Table 12a shows four representative lengths.

At $L = 64$, we also compare the current regularised second-moment source, a Ledoit–Wolf estimate after cross-sample centering, and the current source with $\widehat{m}_{c,L}\widehat{m}_{c,L}^\top$ subtracted, all against the fixed independent reference used above. Their normalised oracle risks agree to five decimal places, while the excess above the reference floor changes from 0.33% to 0.34%, a difference of 0.01 percentage points (Table 12b). The centered covariance source has lower risk in 114 of 263 channels. These measurements quantify the approximation made by retaining the second moment in the implemented source; they do not equate a nonzero-mean second moment with a covariance.

Table 12: **Residual means and covariance centering on HumanML3D.** (a) Four of the 40 measured lengths, using the window law of Eq. (8); median and maximum are over 263 channels. (b) Oracle risk at $L = 64$ against the same independent validation reference as Table 11. Excess is relative to the matched reference floor and is computed before rounding the risk values.

(a) Mean contribution to temporal second moments

L	windows	e_L	$q_{c,L}$ (median / max.)	$\gamma_{c,L}$ (median / max.)
40	24,546	0.24%	0.04% / 3.5%	1.0004 / 1.037
64	22,498	0.30%	0.01% / 4.5%	1.0001 / 1.047
128	14,214	0.85%	0.07% / 7.9%	1.0007 / 1.086
196	8,620	1.77%	0.38% / 7.9%	1.0039 / 1.085

(b) Effect on the independent-reference oracle risk

source or reference	$S_{24}/\sqrt{\operatorname{tr} \widehat{B}_{\mathrm{ref}}}$	excess above floor
matched reference floor	0.05218	–
second moment + LW (current)	0.05235	+0.33%
cross-sample-centered covariance + LW	0.05235	+0.34%
current source minus $\widehat{m}\widehat{m}^{\top}$	0.05235	+0.34%

B.8 THE CROSS BLOCK

For a Toeplitz matrix the cross block is $K = HAu = (s - d\mathbf{1})/\sqrt{L}$ with s the vector of row sums, i.e. the non-uniformity of the row sums near the window edges; adding a constant to the autocorrelation changes d but not K. Measured on training windows (marginal window covariance B_L, cross-sample mean removed, no per-sequence centering, median over channels): $\|HB_Lu\|^2 =$ 9.0, 44, 109 at $L = 64$, 150, 196, against $\|K_{\mathrm{C}}\|^2 = 11.5$, 4.6, 3.7 for the centered Toeplitz and $\|K_{\mathrm{U}}\|^2 = 9.1$, 41, 55 for the uncentered one; the cosine between the data's cross block and K_{C} is 0.94, 0.69, 0.15, and with K_{U} it is 0.95, 0.91, 0.79. A per-sequence constant offset contributes no cross block ($H\mathbf{1} = 0$), so the data's large cross block says that the level is not a constant per clip but a slow trend whose window mean is correlated with the low-frequency shape inside the window. The uncentered Toeplitz reproduces this better than the centered one at every length, and the source with no cross block is farthest from the data in this statistic; the three-seed mean FIDs (uncentered Toeplitz 0.135, its zero-cross-block version 0.139, centered Toeplitz 0.234, its zero-cross-block version 0.222, TSCM-Toeplitz 0.151) do not follow that distance in either direction. In the oracle of Proposition 3, removing the cross block changes the Euler risk by less than 1.5% for either background.

Removing the block changes the three-seed mean test FID by -0.012 on the centered background ($0.234 \to 0.222$) and $+0.004$ on the uncentered one ($0.135 \to 0.139$), differences comparable to the seed-to-seed variation in Table 9.

Validation-selected comparisons. We select each checkpoint by the common validation-FID rule of Appendix B.2 and hold it fixed during the guidance search. Under this protocol, the cross block moves FID by no more than a run's own checkpoint variation, while the level effect is 0.10 and consistent across both backgrounds and the level-rescaled variants; the reference-flow experiment of Appendix B.9 shows that neither is a solver artefact.

B.9 REFERENCE-FLOW EXPERIMENT ON FROZEN MODELS

For a frozen checkpoint and its deployed guidance scale, 192 validation texts are integrated from identical source draws with uniform Euler at $N \in \{12, 24, 48, 96\}$ and with fourth-order Runge–Kutta (RK4) references. Uniform RK4 at 192 and 384 steps agrees to 10^{-3} of the state norm at $w = 1$ and to 1% of the 96-step residual at the deployed scale for the non-white sources. We report the endpoint distance to the reference split into level and fluctuation parts, the FID between the Euler and reference outputs, and each output's FID against the same real motions. These FIDs use only 192 samples and are compared within a model.

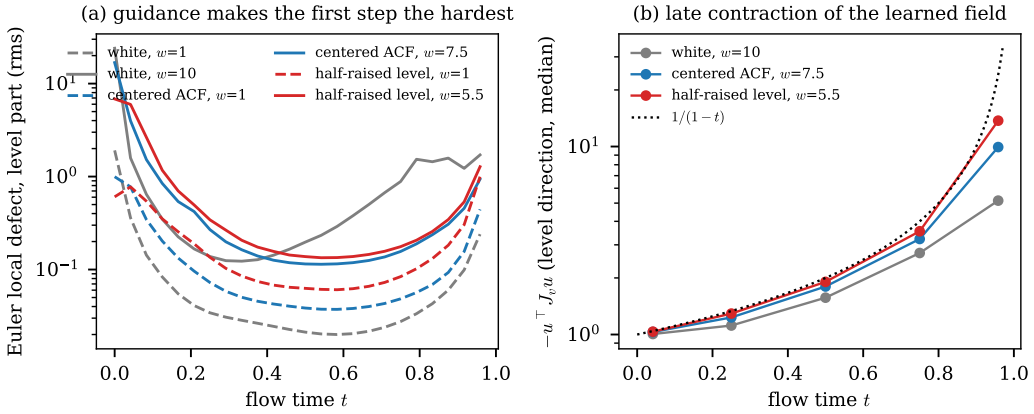

Figure 8: **(a)** Level part of the Euler local defect along the reference trajectory ($h = 1/24$) at $w = 1$ (dashed) and at the deployed guidance scale (solid): guidance makes the first step by far the hardest. **(b)** Signed level-direction Jacobian of the learned guided field against $-1/(1-t)$.

For the stiff white-noise case at $w = 10$, we refine RK4 on a two-part mesh with the fixed breakpoint $t_\star = 1/24$, the end of the first Euler-24 interval. The nodes are

$$t_i = t_\star (i/M)^3, \qquad i = 0, \dots, M,$$
$$t_{M+j} = t_\star + (1 - t_\star) j/T, \quad j = 1, \dots, T.$$

The coarse and fine grids use $(M, T) = (192, 368)$ and $(384, 736)$, respectively, for 560 and 1120 total steps. The cubic first segment concentrates resolution near $t = 0$ (first steps approximately 5.9×10^{-9} and 7.4×10^{-10}); the uniform second segment has step size $1/384$ or $1/768$. The uniform RK4 controls use these same respective step sizes over the entire interval $[0, 1]$. On every interval $[t_i, t_{i+1}]$, standard RK4 with $h_i = t_{i+1} - t_i$ evaluates the guided field at t_i, $t_i + h_i/2$, $t_i + h_i/2$, and $t_i + h_i$. Each field evaluation uses both the conditional and unconditional branches for CFG, and integration is in single precision.

The two graded references have endpoint RMS difference 0.024, less than 0.02% of the Euler-96 endpoint residual of approximately 140. We use the finer graded solution as the endpoint reference. Uniform RK4-384 and RK4-768 differ from it by 13.48 and 0.598, respectively, compared with 27.1 between the original uniform RK4-192 and RK4-384 references. The graded refinement thus resolves the reference discrepancy in this high-guidance case. Its real-motion FID is 0.504, versus 0.500 for the old RK4-384 reference, 0.686 for Euler-24, and 0.484 for Euler-96; FID(Euler-24, reference) is 0.469. With checkpoint, guidance, texts, and source draws fixed, the Euler-24 excess over the refined reference is $0.686 - 0.504 = 0.182 \approx +0.18$ FID. Replacing the old uniform reference changes the real-motion FID by only 0.004. The lower Euler-96 real-motion FID also illustrates that a more accurate solution of the guided flow need not have a lower FID.

For the other sources at their deployed scales, FID(Euler-24, real) versus FID(reference, real) is 0.426 versus 0.420 for centred ACF, 0.449 versus 0.698 for uncentred Toeplitz, and 0.393 versus 0.570 for half-raised level. The Euler-minus-reference FID difference is therefore approximately zero for the centred source and negative for the hot-level sources, whose reference guided flow has higher FID than their 24-step samples. These within-model metric changes do not form an additive FID error decomposition.

The level direction carries 40–60% of the 24-step Euler error in the frozen models tested here (root-mean-square share), less at larger N. Along the reference trajectory, the local Euler defect $d_n = x_{\mathrm{ref}}(t_n + h) - x_{\mathrm{ref}}(t_n) - h\, v_w(x_{\mathrm{ref}}(t_n), t_n)$ at $h = 1/24$ is concentrated in the first two steps for guided fields (level part 17–24 at $n = 0$ against 0.1 in the middle of the trajectory) and is 10–20$\times$ smaller at $w = 1$ (Figure 8a). The signed level-direction Jacobian of the guided field, $u^\top J_v u$ by central differences, follows $-1/(1-t)$ (Figure 8b): the learned conditional field treats the level as determined and contracts onto the conditional manifold.

B.10 FROZEN-FIELD SOURCE AND FREQUENCY DIAGNOSTICS

Across-source reference flows. We freeze the final epoch-400 EMA checkpoint for each of the 16 sources and three training seeds, disable guidance, and pair source draws on the same 1024 HumanML3D test texts. For each checkpoint, let $E_{\mathrm{int}} = \mathrm{rms}\,\|X_{24} - X_{\mathrm{ref}}\|$ be the endpoint residual between Euler-24 and RK4-192 in data coordinates. Table 13 reports source means; Figure 9 shows the three-seed range. Over all sources, the log–log Pearson correlations of S_{24} with E_{int} and reference FID are 0.83 and 0.89; the leave-one-source-out range for the residual correlation is 0.80–0.88. Among the ten Toeplitz variants plus TSCM, the corresponding correlations are 0.78 and 0.10. The residual correlation has a leave-one-source-out range of 0.25–0.87: removing the level $\times\frac{1}{4}$ variant reduces it to 0.25. Thus this group shows an association over its tested mismatch range, but not stable fine-grained ordering of near-matched sources. The SE source has the smallest residual (46.8) but reference FID 7.59, illustrating that numerical ease alone does not determine distributional quality.

Table 13: **Unguided reference-flow control for all 16 sources.** S_{24} is the covariance-only oracle risk; E_{int} is Euler-24 versus RK4-192 endpoint RMS in data coordinates. The next two columns are test FID under the same frozen final-EMA checkpoints ($w = 1$), averaged over three training seeds. The last column is the separately validation-selected guided test FID of Table 8. The first 11 rows are the ten Toeplitz variants plus TSCM.

source	S_{24}	E_{int}	$\mathrm{FID}_{\mathrm{ref}}$	FID_{24}	guided FID
TSCM	0.46	55.3	0.91	1.91	0.108
level halfway	0.48	47.3	0.84	1.68	0.141
level $\times 4$	0.49	49.7	0.94	1.71	0.167
uncentered Toeplitz	0.50	53.9	0.94	1.90	0.135
level $\times 4$, trace-normalized	0.50	53.4	0.86	1.69	0.173
(F,F)	0.50	55.8	0.87	1.84	0.139
TSCM-Toeplitz (T,F)	0.50	53.1	0.90	1.68	0.151
(T,T)	0.54	54.7	0.90	1.60	0.222
(F,T)	0.54	57.2	0.95	1.87	0.218
centered Toeplitz	0.55	51.3	1.08	1.76	0.234
level $\times\frac{1}{4}$	0.69	68.9	0.87	1.42	0.431
TSFlow OU	0.58	55.1	1.01	1.96	0.207
TSFlow SE	0.78	46.8	7.59	9.14	0.485
white noise	1.21	75.2	2.36	4.14	0.497
time-permuted Toeplitz	1.89	91.9	4.69	11.48	0.554
TSFlow periodic	5.11	84.9	16.81	22.28	1.867

One common guidance scale. For the same 11-source group, we hold one final-EMA checkpoint per source fixed, use the same test texts and base noise draws across sources, and scan common guidance strengths at 24 Euler steps. Figure 10 shows that the S_{24}–FID correlation is negative at $w = 1$ (Pearson -0.79), positive at $w = 3$ and 5.5 (0.80 and 0.84), and negative again at $w = 10$ (-0.17). At $w = 5.5$, Spearman correlation is 0.67 and leave-one-source-out Pearson correlations span 0.78–0.86. The positive association in this moderate-guidance range does not require selecting a different w for each source; the dotted line shows the 0.96 correlation from the separately validation-selected checkpoint and scale in Table 8. This scan does not separate continuous-time guidance bias from a guidance–discretization interaction, since it does not include guided reference integrations.

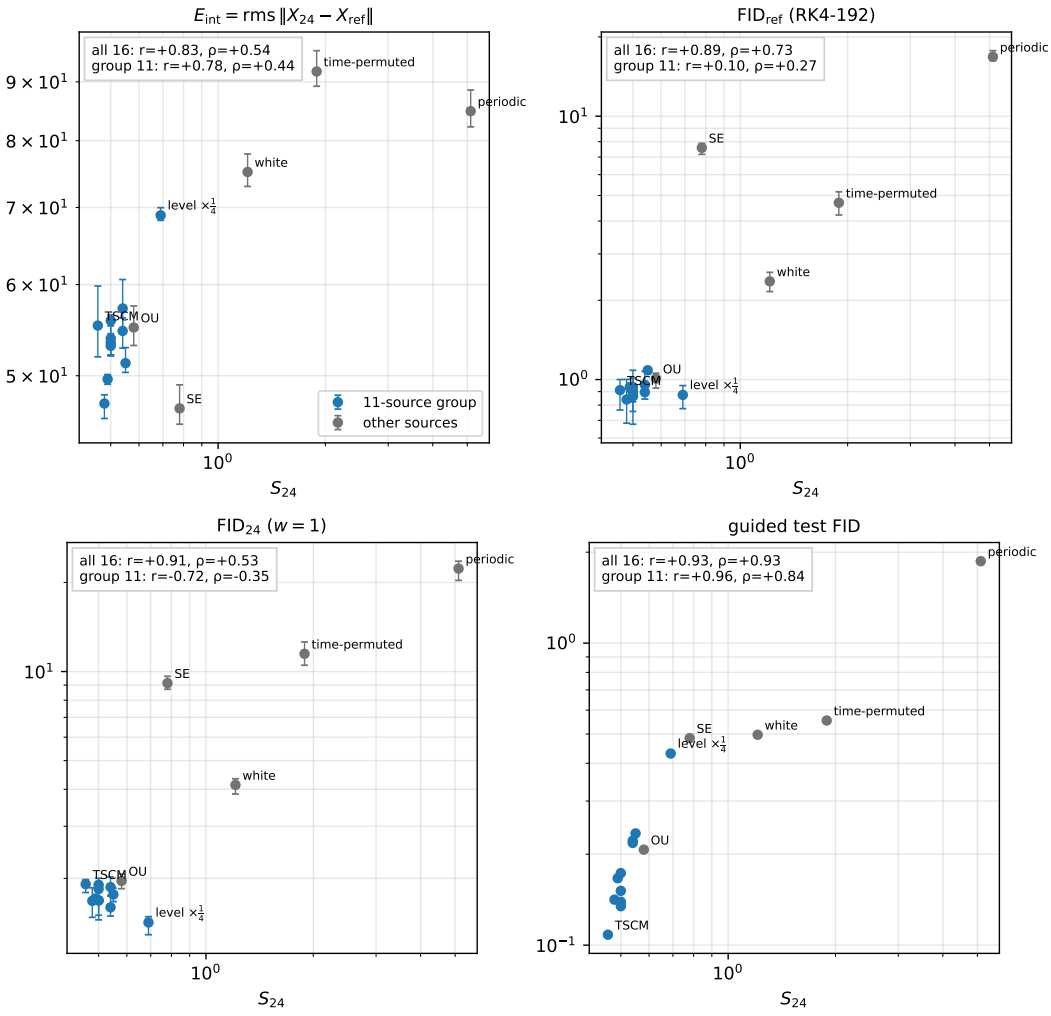

Figure 9: **Oracle risk versus frozen-field outcomes.** The upper panels show the unguided Euler-to-RK4 residual and RK4-192 FID; the lower panels show unguided Euler-24 FID and the separately selected guided test FID of Table 8. Points are source means over three training seeds; bars on unguided panels span their seed minima and maxima. Blue marks the ten Toeplitz variants plus TSCM, gray the other five sources. In-panel r and ρ are log–log Pearson and Spearman correlations, respectively.

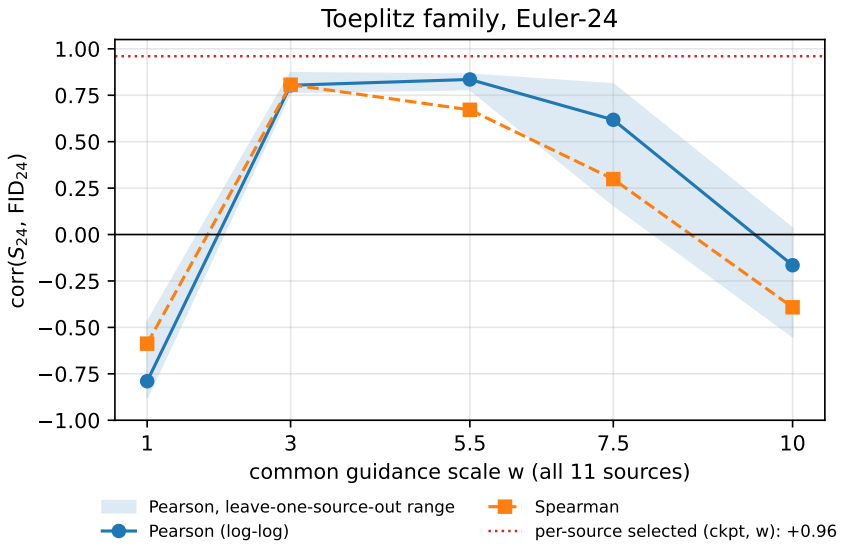

Figure 10: **Fixed-guidance control for ten Toeplitz variants plus TSCM.** One fixed final-EMA checkpoint per source is evaluated at each common guidance strength w; all sources share test texts and base noise draws. Solid: log–log Pearson correlation between S_{24} and Euler-24 FID; dashed: Spearman correlation; shaded: leave-one-source-out Pearson range. The dotted line is the separately validation-selected checkpoint-and-scale result from Table 8.

White/TSCM frequency diagnostics. We turn guidance off and examine all six final epoch-400 EMA checkpoints of the white-noise and TSCM DiT-M arms (three training seeds per source; no retraining). For each checkpoint, 1024 test texts are sampled with paired random seeds using 24 Euler steps and an RK4-192 reference. On a 192-sample subset, the relative RMS change from RK4-192 to RK4-384 is at most 7.7×10^{-4} for white noise and 5×10^{-6} for TSCM. Against the same real motions, the reference FIDs are 2.36 ± 0.20 and 0.91 ± 0.13 (mean $\pm$ SD across seeds); the Euler-24 FIDs are 4.14 ± 0.25 and 1.91 ± 0.10. Thus a learned-flow gap remains after integration is refined. The within-model Euler-minus-reference FID changes are 1.78 ± 0.07 and 1.00 ± 0.03, respectively; FID is nonlinear, so these changes are not an additive decomposition of the cross-model gap.

We Fourier-transform each channel over its valid L frames and measure $|X_k|^2/L$ in the DC, $(0, 1/12]$, $(1/12, 1/4]$, and $(1/4, 1/2]$ cycles-per-frame bands. The converged white and TSCM flows retain, respectively, 0.56 ± 0.01 and 0.75 ± 0.02 of the data's per-channel variance. Their generated-to-real power ratios are $0.68/0.57/0.31$ and $0.81/0.70/0.51$ in the DC/low/mid bands. The high-band aggregate instead gives 0.84 versus 0.76; within the smooth joint-position and rotation channels, white noise gives 1.37 and 1.49 versus TSCM's 0.98 and 0.82. Figure 5 also shows the paired Euler-minus-reference endpoint power. These Fourier projections locate differences in learned outputs and solver error, but need not coincide with the generalized eigenmodes of Proposition 1 or explain FID by themselves.

As a second diagnostic, consider one Fourier coefficient with independent source and target second moments a and b. The optimal predictor of its flow-matching target that is linear in the same coefficient of x_t has population squared error

$$L_{\mathrm{lin}}(t; a, b) = \frac{ab}{(1-t)^2 a + t^2 b}.$$

The network also sees other coefficients and the text. Hence its population optimal loss is at most L_{lin}, and a positive population gap between the network loss and this restricted benchmark lower-bounds its squared error relative to the optimal conditional field. This is a linear benchmark, not an irreducible error floor. For each channel and frequency bin, we plug sample-mean source and real powers into L_{lin} and clip the estimated gap at zero before pooling within a band. On 512 validation windows at ten times from 0.05 to 0.95, the positive plug-in gap at $t = 0.05$ in the high band is

39.9% of that band's data power for white noise and 10.4% for TSCM. At $t = 0.95$, 82% versus 8% of DC (frequency-bin, channel) cells exceed the benchmark. Figure 11 shows the full time grid. These finite-sample gaps are descriptive; clipping each estimated cell gap at zero can increase their reported average.

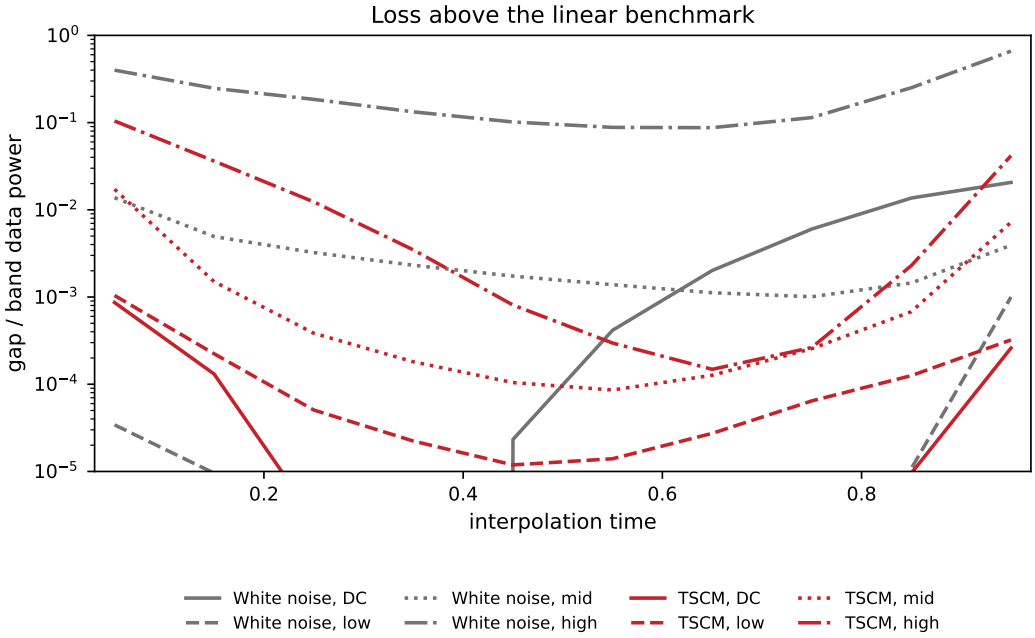

Figure 11: **Frequency-resolved field-fitting gap for unguided DiT-M.** Estimated positive loss above the restricted linear benchmark, divided by each band's data power, on 512 validation windows. Gray is white noise and red is TSCM; solid, dashed, dotted, and dash-dotted lines denote DC, low, mid, and high bands. The full legend is shown below the panel. This is the companion diagnostic to Figure 5; finite-sample clipping makes the curves descriptive.

B.11 Higher-order solvers at equal cost

The reference-flow experiment above diagnoses finite-step integration error; here we test whether the TSCM advantage persists with higher-order solvers at the same sampling budget. We use one frozen checkpoint for each of the white-noise, centered-Toeplitz and TSCM source designs of Table 8, keeping the checkpoint fixed across solvers. We compare 24 uniform Euler steps, 12 uniform Heun steps and six uniform classical Runge–Kutta (RK4) steps. Each uses 24 guided-field evaluations, with conditional and unconditional passes combined at every stage exactly as for Euler. The Euler column retains the stored configuration selected on the original validation guidance grid. For Heun and RK4, the constant guidance scale is selected separately on validation from neighbouring candidates on the model's guidance ladder.

Table 14 reports test FID over 20 sampling repeats for each configuration. Heun does not improve the white-noise or centered-Toeplitz model. RK4 reduces their FID by 28% and 7%, respectively, consistent with the greater numerical burden of the white-noise source in the preceding diagnostic. TSCM reaches 0.087–0.090 under the higher-order solvers, both selected at $w = 4$; these comparisons include the effect of guidance re-tuning as well as the change of integrator. Across Euler, Heun and RK4, white-noise FID remains 4.1, 5.3 and 3.9 times TSCM's, respectively. Even with RK4, the white-noise model does not reach the centered-Toeplitz model's Euler-24 FID. Thus, on these frozen models, TSCM's advantage persists when Euler is replaced by higher-order solvers at equal guided-field evaluation cost and guidance is re-tuned for each replacement solver.

Table 14: **Solvers at equal cost (24 guided-field evaluations).** HumanML3D test FID (mean ± SD over 20 sampling repeats) for one frozen checkpoint per source, with the same checkpoint used across columns. w is the validation-selected constant guidance scale; Euler retains the stored selection from the original guidance grid, while Heun and RK4 are tuned separately.

model	Euler-24	Heun-12	RK4-6
white noise	.471 ± .014 (w=10)	.480 ± .011 (w=7.5)	.337 ± .012 (w=7.5)
centered Toeplitz	.205 ± .010 (w=7.5)	.214 ± .010 (w=7.5)	.190 ± .008 (w=5.5)
TSCM	.116 ± .007 (w=5.5)	.090 ± .005 (w=4)	.087 ± .005 (w=4)

B.12 GUIDANCE TRAJECTORIES

The frozen TSCM checkpoint (selected at $w = 5.5$) is integrated with 24 Euler steps from the same source draws (192 validation texts, four draws each) under $w \equiv 1$, $w \equiv 5.5$, $w \equiv 7.5$, and the early-window schedules. At each step we evaluate the conditional and unconditional fields, split $\Delta = v_c - v_u$ into its level part (per-channel mean over valid frames, scaled by $\sqrt{L}$) and fluctuation part, and record the cosine between the level part at step n and at step 0. At $t = 1$ we compute per-text, per-channel levels of the generated motions and of the real motion of the same text; the between-text variance is the variance over texts of the mean over the four draws, the within-text variance is the variance over draws (0.16–0.18 for every policy; guidance changes the spread between texts). Measured values: $\|\Delta_{\text{level}}\|$ drops from 82 to within a factor two of its plateau (about 11) in four steps under guidance and seven without, and its direction decorrelates from the $t = 0$ direction (cosine to the initial direction 0.55 at step 3 without guidance, 0.19 with $w = 5.5$); the endpoint level variance is 0.39 ($w = 1$), 0.78 ($w = 5.5$), 0.91 ($w = 7.5$), 0.66 (the early-window schedule $3 \to 5.5$) and 0.60 ($2 \to 7.5$) against 0.54 for the real motions of the same texts; the policies of Figure 4c differ only in the first two steps, and the late field's level-direction Jacobian follows $-1/(1-t)$ (Appendix B.9), a contraction of each sample onto its conditional manifold that does not undo a population-wide shift. The sub-step control (on the TSCM-Toeplitz model) integrates the first $M \in \{1, 2, 3\}$ intervals with $K \in \{8, 8, 4\}$ Euler sub-steps at constant $w = 5.5$ and evaluates the full validation split once: 0.136 (plain), 0.177, 0.194, 0.186.

Covariance-bias measurement (Figure 4b). On the same checkpoint, 384 validation texts with eight source draws each are integrated with RK4 at step $1/192$ from $t = 0$ under constant $w \in \{1, 2, 3, 4, 5.5, 7.5, 10\}$ from identical (x_0, y) pairs, in single precision. At every step the level $u^\top X_t$ is recorded per channel and $\Delta\mathrm{Var}(t; w) = \mathrm{Var}[u^\top X_t^{(w)}] - \mathrm{Var}[u^\top X_t^{(1)}]$ (paired, averaged over channels) is compared with $t^2 w(w-1)\, u^\top V u$, where $u^\top V u$ is the between-text variance of the level of $\Delta(x_0, 0)$ (repeat mean per text, within-text part removed), i.e. $\mathrm{Var}_Y(u^\top \mu_Y)$ by the initial-translation identity in Appendix A.13.1. Measured over predicted is 1.0–1.3 for $t \le 0.02$ (0.9–1.0 on the 96 full-length texts alone at $t = 0.016$–0.021, where that smaller estimate is above its sampling noise) and decreases to 0.2–0.65 at $t = 0.125$.

B.13 EARLY-WINDOW GUIDANCE: PAIRS AND VARIANTS

Table 15 gives the with/without pairs of Section 4 for the sources of the controlled study on which the schedule was swept and for the latent model. Under the schedule the step-count optimum moves: on validation TSCM-Toeplitz is best at 48–64 steps (0.087 against 0.095 at 24) and TSCM at 24 (0.102); Tables 15 and 16 use 24 steps. Table 16 lists variants tested on the centered-Toeplitz DiT-M checkpoint (validation, one repeat; constant $w = 7.5$ gives 0.218). In this sweep, retaining weak early guidance outperforms setting $w_0 = 1$; a longer weak-guidance window, a slow ramp, and guidance restricted to $t \in [0.1, 0.9]$ also perform worse than the best tested schedule. Reducing guidance after $t \ge 0.75$ gives no improvement over constant 7.5. These observations concern this checkpoint and tested schedules, rather than a universal restriction on w_0: the MDM experiment in Appendix B.15 uses $w_0 = 1$ successfully.

Table 15: **Early-window guidance.** Single-training-run with/without pairs for the sources of Table 8 and the latent T/4 model of Section B.16, HumanML3D test FID (mean ± SD over 20 repeats, 24 steps; "constant" is each model's swept scale, the schedule is selected on validation; TSCM uses its own best constant scale as w_1, the Toeplitz sources the next candidate above). Variants on the centered-ACF model are in Table 16.

model	white noise	centered Toeplitz	centered, level ×4	swap (F,T)	swap (F,F)	half-raised level	TSCM-Toeplitz	**TSCM**	latent T/4, TSCM
constant w	.471±.014	.205±.010	.168±.006	.233±.009	.135±.008	.141±.009	.131±.007	.116±.007	.097±.006
(w_0, steps, w_1)	(3,1,11.5)	(3,3,10)	(2,2,7.5)	(3,3,10)	(3,4,7.5)	(3,4,7.5)	(2,3,7.5)	(3,2,5.5)	(2,4,4)
with schedule	.221±.006	.090±.007	.088±.006	.096±.007	.083±.006	.087±.008	.079±.005	**.076±.005**	.083±.005

Table 16: **Schedule variants on the centered Toeplitz model** (validation FID, one repeat).

variant	FID	variant	FID
2 steps at $w = 2$ / 2.5 / 3, then 7.5	.149 / .152 / .156	3 / 6 steps at $w = 1$, then 7.5	.305 / .815
3 steps at $w = 3$, then 10	.124	ramp $1 \to 7.5$ over $t \in [0, 0.5]$	.409
4 steps at $w = 3$, then 7.5	.224	guidance only for $t \in [0.1, 0.9]$	.304
constant 5.5 / 10	.297 / .277	constant 7.5, reduced after $t \geq 0.75$	.224–.236

B.14 FINEDANCE

FlowerDance (Yang et al., 2025) is a released music-to-dance model on FineDance (Li et al., 2023). We continue the same checkpoint for 10 epochs with its recipe and no guidance, changing only the source. TSCM replaces the prior on contact and joint-rotation channels, retains white noise on absolute root translation, and uses length-indexed statistics for the 240-frame training slices and 1020-frame test sequences. Under the FineDance protocol, TSCM lowers FID_k from 29.75 to 21.94 (Table 17).

Table 17: **Music-to-dance, FineDance test set.** FlowerDance fine-tunes with the same released checkpoint and 10-epoch budget, differing only in source. ΔFID_k is relative to the white-prior fine-tune.

method	$\text{FID}_k \downarrow$	ΔFID_k	$\text{Div}_k \uparrow$	$\text{Div}_g \uparrow$	BAS↑
real motions	–		9.94	7.54	0.201
FlowerDance +10 epochs, white prior	29.75	+0%	8.14	7.38	0.232
+10 epochs, TSCM prior	**21.94**	−26%	**8.99**	**7.44**	0.232

B.15 DIFFUSION MODELS: x_0-PREDICTION DIT-M AND MDM

The analysis concerns the source, not the parameterisation. We trained the same DiT-M as an x_0-prediction diffusion model (1000 linear-β noise levels, UniPC (Zhao et al., 2023) with 10 steps, i.e. the Light-T2M (Zeng et al., 2025) recipe on our backbone) with white noise and TSCM. The test FID improves from 0.173 ± 0.003 to 0.108 ± 0.003 (-38%, 95% confidence intervals over 20 repeats; Table 7), and the validation step-count curves cross at about 40 steps: the source helps at few steps and not beyond in this diffusion experiment. Plain ϵ-prediction with the same solver collapses for both sources in our runs.

MDM. We trained the official MDM (Tevet et al., 2023) (x_0-prediction diffusion, 50-step DDPM sampling, EMA as in its released 50-step recipe) on HumanML3D with TSCM as the only change. Our matched white-source rerun has test FID 0.537 ± 0.044; replacing the source gives 0.380 ± 0.005 at the swept constant scale ($w = 6$) and 0.205 ± 0.004 with the early-window schedule ($w_0 = 1$ for the first 10% of the path, then $w_1 = 7$). These are 95% confidence intervals over 20 test generations; the latter schedule also changes guidance.

Light-T2M. The white-source and TSCM results in Table 3 are our own trainings of Light-T2M (Zeng et al., 2025) using its official standard implementation and settings, with matched budgets and evaluation. Their constant-guidance test FIDs are 0.042 ± 0.002 and 0.035 ± 0.001 (95%

confidence intervals over 20 test generations). The early-window row additionally changes the inference guidance schedule. These trained-model results differ in provenance from the downloaded-checkpoint retest in Table 18 and the published number reproduced in Table 19.

Selection of additional motion models. We attempted TSCM transfers to newer published architectures. Every white-source reproduction used its authors' released standard implementation and training and evaluation settings. Before inspecting the rerun results, we set the reporting rule: an external-backbone pair is included only when its matched white-source rerun differs from the published FID by at most three times the half-width of the rerun's 20-generation 95% confidence interval (e.g., 0.537 ± 0.044 gives a tolerance of 0.132). The reported external-backbone transfers are MDM and Light-T2M; results outside this screen are not included. As a separate split diagnostic, Table 18 reports local retests of downloaded MotionHiFlow and Light-T2M checkpoints and one of our own with the official evaluator. At matched set size, the MotionHiFlow checkpoint's two test-subset FIDs remain roughly half its validation FID. The cause is not identified here; these checkpoint scores are not TSCM transfer comparisons.

Table 18: **Local checkpoint retests across HumanML3D evaluation splits.** The MotionHiFlow and Light-T2M rows use downloaded released weights; E1 is our checkpoint. All rows use the official evaluator on our machine and report means over 20 generation repeats per split. The two random test subsets each match the validation-set size. E1 uses early guidance $(w_0, T_0, w_1) = (2, 0.1, 7.5)$.

model and guidance	full test	test subset 1	test subset 2	validation
MotionHiFlow (Li et al., 2026), released, $w = 4.5$	0.0318	0.0575	0.0599	0.120
Light-T2M (Zeng et al., 2025), released, $w = 4$	0.039	0.0738	0.0755	0.0835
E1 (ours), early guidance	0.0765	0.105	0.114	0.0946

Published text-to-motion comparison. Table 19 places our DiT-L variants alongside the published comparison of Li et al. (2026). Every prior-method number, including Light-T2M's HumanML3D FID 0.040, is transcribed from Table 1 of that paper rather than re-evaluated here; the six DiT-L rows are our results from Table 3. The Light-T2M released-checkpoint FID 0.039 in Table 18 is our local retest, while the 0.042 white-source FID in Table 3 is our separate retraining with the official recipe. The cited paper gives the original reference for each baseline method. Our HumanML3D TSCM+EW and the published MotionHiFlow result both round to FID 0.032, with reported 95% confidence-interval half-widths 0.001 and 0.002, respectively.

Table 19: **Published text-to-motion comparison with our DiT-L variants.** Published-method values are from Table 1 of Li et al. (2026); our DiT-L values are from Table 3. We reproduce its R-precision, FID, and diversity columns. $\pm$ denotes the reported 95% confidence-interval half-width. Bold and underline indicate the best and second-best point estimates in each dataset and metric; both HumanML3D FIDs rounding to 0.032 are marked best. Diversity is shown without ranking.

method	venue	R-precision↑			FID↓	Diversity→
		top-1	top-2	top-3		
HumanML3D						
TEMOS	ECCV'22	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$8.973^{\pm.071}$
T2M-GPT	CVPR'23	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$9.761^{\pm.081}$
ReMoDiffuse	ICCV'23	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$9.018^{\pm.075}$
MoMask	CVPR'24	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	–
BAMM	ECCV'24	$0.525^{\pm.002}$	$0.720^{\pm.003}$	$0.814^{\pm.003}$	$0.055^{\pm.002}$	$9.717^{\pm.089}$
MoGenTS	NeurIPS'24	$0.529^{\pm.003}$	$0.719^{\pm.002}$	$0.812^{\pm.002}$	$\underline{0.033}^{\pm.001}$	$9.570^{\pm.077}$
Light-T2M	AAAI'25	$0.511^{\pm.003}$	$0.699^{\pm.002}$	$0.795^{\pm.002}$	$0.040^{\pm.002}$	–
IRG-MotionLLM	arXiv'25	$0.535^{\pm.002}$	$0.725^{\pm.002}$	$0.820^{\pm.002}$	$0.242^{\pm.006}$	$9.900^{\pm.094}$
EnergyMoGen	CVPR'25	$0.526^{\pm.003}$	$0.718^{\pm.003}$	$0.815^{\pm.002}$	$0.176^{\pm.006}$	$9.500^{\pm.091}$
SALAD	CVPR'25	$\mathbf{0.581}^{\pm.003}$	$\mathbf{0.769}^{\pm.003}$	$\mathbf{0.857}^{\pm.002}$	$0.076^{\pm.002}$	$9.696^{\pm.096}$
MoMask++	NeurIPS'25	$0.528^{\pm.003}$	$0.718^{\pm.003}$	$0.811^{\pm.002}$	$0.072^{\pm.003}$	–
MotionHiFlow	CVPR'26	$0.563^{\pm.003}$	$0.754^{\pm.003}$	$0.843^{\pm.003}$	$\mathbf{0.032}^{\pm.002}$	$9.504^{\pm.071}$
DiT-L, white	ours	$0.558^{\pm.002}$	$0.753^{\pm.002}$	$0.844^{\pm.002}$	$0.345^{\pm.004}$	$9.991^{\pm.085}$
DiT-L, TSCM	ours	$\underline{0.568}^{\pm.002}$	$\underline{0.759}^{\pm.001}$	$\underline{0.847}^{\pm.002}$	$0.047^{\pm.001}$	$9.656^{\pm.096}$
DiT-L, TSCM+EW	ours	$0.562^{\pm.002}$	$0.755^{\pm.002}$	$0.844^{\pm.002}$	$\mathbf{0.032}^{\pm.001}$	$9.553^{\pm.102}$
KIT-ML						
TEMOS	ECCV'22	$0.353^{\pm.006}$	$0.561^{\pm.007}$	$0.687^{\pm.005}$	$3.717^{\pm.051}$	$10.84^{\pm.100}$
T2M-GPT	CVPR'23	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$10.86^{\pm.094}$
ReMoDiffuse	ICCV'23	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$0.155^{\pm.006}$	$10.80^{\pm.105}$
MoMask	CVPR'24	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$0.204^{\pm.011}$	–
BAMM	ECCV'24	$0.438^{\pm.009}$	$0.661^{\pm.009}$	$0.788^{\pm.005}$	$0.183^{\pm.013}$	$11.008^{\pm.094}$
MoGenTS	NeurIPS'24	$0.445^{\pm.006}$	$0.671^{\pm.006}$	$0.797^{\pm.005}$	$0.143^{\pm.004}$	$10.918^{\pm.090}$
Light-T2M	AAAI'25	$0.444^{\pm.006}$	$0.670^{\pm.007}$	$0.794^{\pm.005}$	$0.161^{\pm.009}$	–
IRG-MotionLLM	arXiv'25	$0.445^{\pm.005}$	$0.681^{\pm.003}$	$0.781^{\pm.004}$	$0.432^{\pm.013}$	$11.115^{\pm.086}$
EnergyMoGen	CVPR'25	$0.436^{\pm.006}$	$0.651^{\pm.006}$	$0.772^{\pm.006}$	$0.495^{\pm.020}$	$11.06^{\pm.101}$
SALAD	CVPR'25	$0.477^{\pm.006}$	$\underline{0.711}^{\pm.005}$	$0.828^{\pm.005}$	$0.296^{\pm.012}$	$11.097^{\pm.095}$
MotionHiFlow	CVPR'26	$\underline{0.482}^{\pm.005}$	$0.704^{\pm.005}$	$0.825^{\pm.005}$	$0.135^{\pm.007}$	$10.894^{\pm.117}$
DiT-L, white	ours	$\mathbf{0.488}^{\pm.005}$	$\mathbf{0.728}^{\pm.004}$	$\mathbf{0.844}^{\pm.004}$	$0.172^{\pm.003}$	$11.288^{\pm.122}$
DiT-L, TSCM	ours	$0.475^{\pm.004}$	$0.700^{\pm.004}$	$0.829^{\pm.002}$	$\underline{0.120}^{\pm.002}$	$11.184^{\pm.134}$
DiT-L, TSCM+EW	ours	$0.475^{\pm.004}$	$0.702^{\pm.004}$	$\underline{0.830}^{\pm.002}$	$\mathbf{0.106}^{\pm.002}$	$11.125^{\pm.135}$

B.16 LATENT MODEL

Raw motion is ill-conditioned along two axes. Within a frame, the 263 channels lie on a thin manifold: the condition number of the per-frame covariance at noise level 0.02 is 7×10^4. Along time, the temporal covariance under white noise has condition number 6×10^4 (Section 2). A latent autoencoder addresses the first axis. MoRAE (Zhu et al., 2026) showed that a compact, near-isotropic, full-rank latent removes the within-frame ill-conditioning that otherwise forces the flow to spend capacity on frame validity; our 32-channel, $4\times$ temporally downsampled latent of this kind has a per-frame condition number of 37 (reconstruction FID 0.002, Appendix B). Temporal anisotropy remains: the latent's temporal covariance under white noise still has condition number 10^2–4×10^2, and a Toeplitz source of the latent's centered autocorrelation brings it to 9–64.

B.17 SKY TIME-LAPSE WITH WAN2.1

Wan2.1-T2V-1.3B (Wan Team, 2025) is a rectified-flow text-to-video DiT with a $4 \times 8 \times 8$ causal 3D VAE (16 latent channels) and native classifier-free guidance. Data: the Sky Time-lapse dataset (Xiong et al., 2018); from each of the 994 training videos we cut two non-overlapping 33-frame windows of consecutive frames and from the test videos 512 windows, centre-cropped and resized to 256×256. Every clip is encoded once by the VAE to a $16 \times 9 \times 32 \times 32$ latent. The text condition is a single fixed caption ("A time-lapse video of the sky with clouds drifting and changing shape.") encoded once by the umT5 encoder; the negative prompt of the released code serves as the unconditional context.

The TSCM table is the 9×9 uncentred temporal second moment of each latent channel, pooled over all training clips and spatial positions, with Ledoit–Wolf shrinkage and a Cholesky factor; the source draw applies the factor along the latent time axis and is white across space and channels. Because the sky is nearly static, the level carries 93% of the latent second-moment power (median $u^\top A u = 5.6$ against 1 for white noise, of a trace of 6).

Fine-tuning: all weights, rectified-flow velocity loss $\|v_\theta(x_t, t) - (z - x_1)\|^2$ with $x_t = (1-s)x_1 + sz$ and the shifted timestep density of the released sampler (shift 3), condition dropout 0.1, AdamW (10^{-5}, $\beta = (0.9, 0.95)$, 50 warm-up steps), batch 4, three epochs (1,128 updates), seed 3407, bf16 autocast with fp32 master weights on one GPU. The two arms differ only in the distribution of z: white noise or TSCM. No selection is made: the last checkpoint of each arm is evaluated. Sampling: UniPC with shift 3 and guidance 5 (the released defaults), the same 512 initial-noise seeds for every arm, the initial noise drawn from the arm's own source. FVD: I3D features (the Kinetics-400 network of TATS, via cd-fvd) on the 16 frames $0, 2, \ldots, 30$ of each clip, against the 512 real test clips.

B.18 TSFlow unconditional generation: details and 16-step results

TSFlow (Kollovieh et al., 2025) trains an unconditional flow-matching model per dataset from a Gaussian-process source (an OU, squared-exponential or periodic kernel) and scores it by the 2-Wasserstein distance between generated and real windows after a few Euler steps. We follow the unconditional-generation task and evaluation code of Table 1 in Kollovieh et al. (2025), including its training-window reference distribution, and change only the source. Every result is our re-run of the same released code with the same backbone, training recipe, hyper-parameters, and evaluation procedure on the same eight datasets. The setting uses an S4 backbone, independent coupling, and windows of 240–384 steps; the sources are white noise, the three released kernels, TSCM-Toeplitz, and TSCM (the second moment of 2,000 random windows per series in the model's coordinates, one $W \times W$ table per dataset).

Setting of Kollovieh et al. (2025), unconditional models: S4 backbone (3 layers, width 64), learning rate 10^{-3}, gradient clipping 0.5, batch 64, 128 batches per epoch, 1,000 epochs, EMA 0.9999, independent coupling, one model per dataset and source; windows are context plus prediction length ($336 + 24$ for the hourly sets, $336 + 48$ for KDD-Cup and M4-hourly, $210 + 30$ for the daily sets), in the model's coordinates ($x = \mathrm{window}/\bar{x}_{\mathrm{train}} - \mathbf{1}$, where $\bar{x}_{\mathrm{train}}$ is the series' training mean). The subtraction of 1 is applied before flow-matching interpolation and is included in the coordinates used to estimate the source. The kernels use the released code and its unit path noise; TSCM and TSCM-Toeplitz are estimated from 2,000 random windows per training series (at most 100,000), TSCM as the shrunk $W \times W$ second moment (Ledoit–Wolf, noise floor $10^{-3}\,\mathrm{tr}\,/W$), TSCM-Toeplitz as the construction of Section 3.2 with unnormalised autocovariances so that its variance matches the data. The score is the exact 2-Wasserstein distance (squared Euclidean cost) between 10,000 generated windows and 10,000 random training windows, from the final EMA checkpoint, at 4 and 16 uniform Euler steps; three training seeds per cell.

Residual-mean check. In these shifted model coordinates, the mean-energy ratio $\|\widehat{m}\|^2 / \mathrm{tr}\,\widehat{S}$ is at most 0.83% (Exchange); the other seven datasets lie between 2×10^{-5} and 10^{-3}. Across the eight datasets, the mean of the temporal window mean ranges from -0.023 to 0.026, and the level ratio $u^\top \widehat{S} u / (u^\top \widehat{C} u)$ is at most 1.009, with $\widehat{C} = \widehat{S} - \widehat{m}\widehat{m}^\top$ as in Table 12. For the same-window oracle comparison, B is the estimated covariance with a 10^{-3} noise floor and $A_{\mathrm{cov}} = B$, so $S_N(A_{\mathrm{cov}}; B)/\sqrt{\mathrm{tr}\,B} = 1 - R_N(1)$ is the matched reference floor. In the model coordinates, the relative oracle risk increase of the second-moment source, $S_N(A_{\mathrm{second}}; B)/S_N(A_{\mathrm{cov}}; B) - 1$, is at most 2×10^{-5} across all eight datasets and $N \in \{4, 16\}$. The maxima are 1.82×10^{-5} and 1.90×10^{-5}, respectively, both on Exchange.

B.19 Illustrative samples

These selected samples illustrate outputs and evaluation settings. Quantitative claims rely on the matched distributional evaluations and numerical diagnostics; individual renderings do not establish distributional or perceptual superiority. Motion panels overlay poses from one sequence on a shared

Table 20: **The unconditional-generation task of Table 1 in Kollovieh et al. (2025)**, W_2 at 16 Euler steps; every row our re-run, mean $\pm$ SD over three training seeds.

dataset	white noise	TSFlow (OU)	TSFlow (SE)	TSFlow (periodic)	TSCM-Toeplitz	TSCM
Electricity	6.21 ± 0.41	6.02 ± 0.14	6.15 ± 0.21	5.74 ± 0.05	6.12 ± 0.30	$\mathbf{5.72} \pm 0.15$
Solar	12.50 ± 0.31	12.92 ± 1.01	12.53 ± 0.27	10.74 ± 0.19	$\mathbf{10.55} \pm 0.82$	11.12 ± 0.10
Traffic	10.91 ± 0.29	11.38 ± 0.38	11.20 ± 0.13	10.99 ± 0.23	10.82 ± 0.03	$\mathbf{10.51} \pm 0.02$
UberTLC	45.93 ± 2.13	44.74 ± 1.70	45.39 ± 6.95	41.92 ± 1.83	45.68 ± 4.32	$\mathbf{41.49} \pm 2.51$
KDD-Cup	13.60 ± 0.02	13.46 ± 0.12	13.47 ± 0.11	13.54 ± 0.19	13.72 ± 0.10	$\mathbf{13.37} \pm 0.22$
M4-hourly	7.12 ± 0.26	7.23 ± 0.24	7.58 ± 0.37	7.67 ± 1.04	7.81 ± 0.42	$\mathbf{6.48} \pm 0.08$
Exchange	0.82 ± 0.24	0.57 ± 0.03	0.61 ± 0.06	0.56 ± 0.01	0.61 ± 0.12	$\mathbf{0.54} \pm 0.09$
Wiki2000	32.11 ± 0.12	31.57 ± 0.02	31.83 ± 0.02	31.80 ± 0.21	31.57 ± 0.12	$\mathbf{31.55} \pm 0.40$

ground plane, while dance and video panels show frames in temporal order. Each caption specifies which model, source, training, or guidance settings differ.

Text-to-motion: examples from different generators.

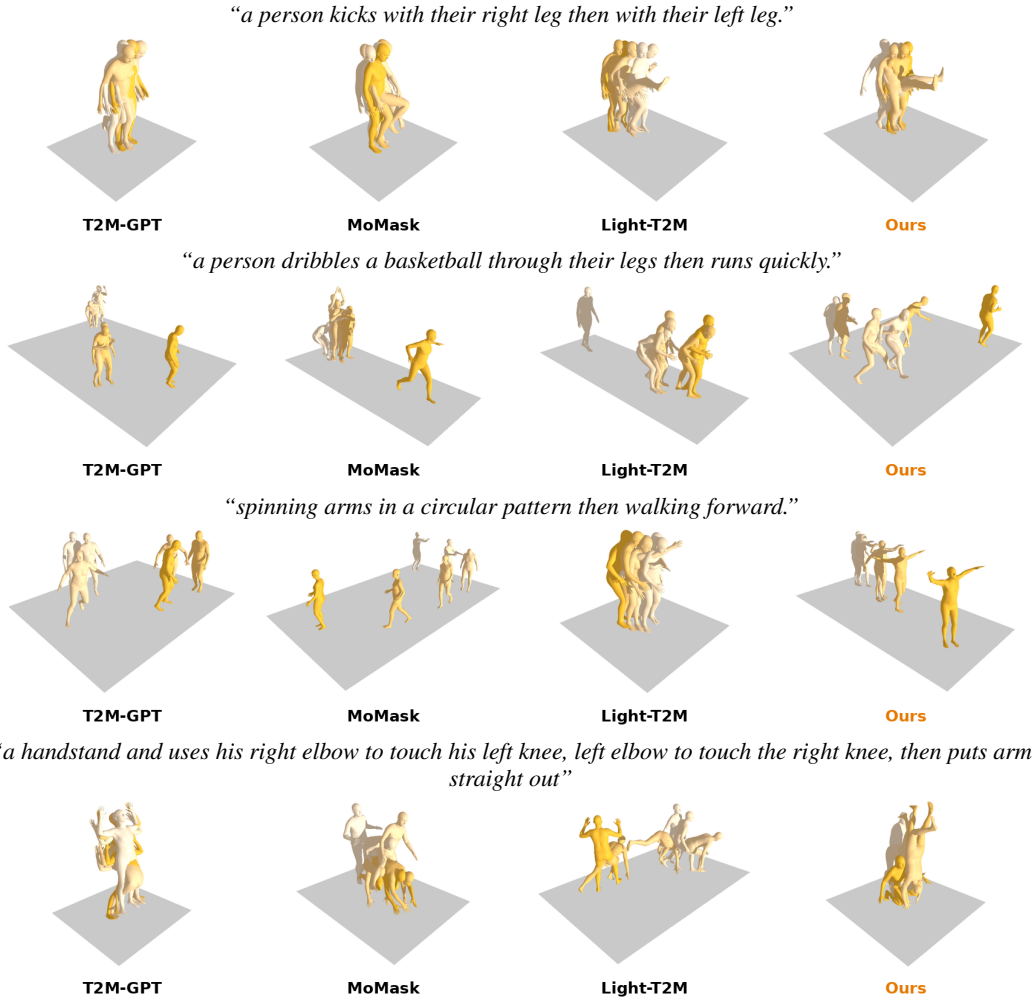

Figure 12: **Text-to-motion examples under the same captions.** Columns show T2M-GPT (Zhang et al., 2023), MoMask (Guo et al., 2024), Light-T2M (Zeng et al., 2025), and the TSCM generator (orange label). Each panel overlays poses from one generated sequence; objects mentioned in the text are not rendered. Architectures and training recipes differ across columns, so this panel does not isolate source replacement. The corresponding within-model quantitative comparisons are in Table 3.

Text-to-motion: source, conditioning, and guidance.

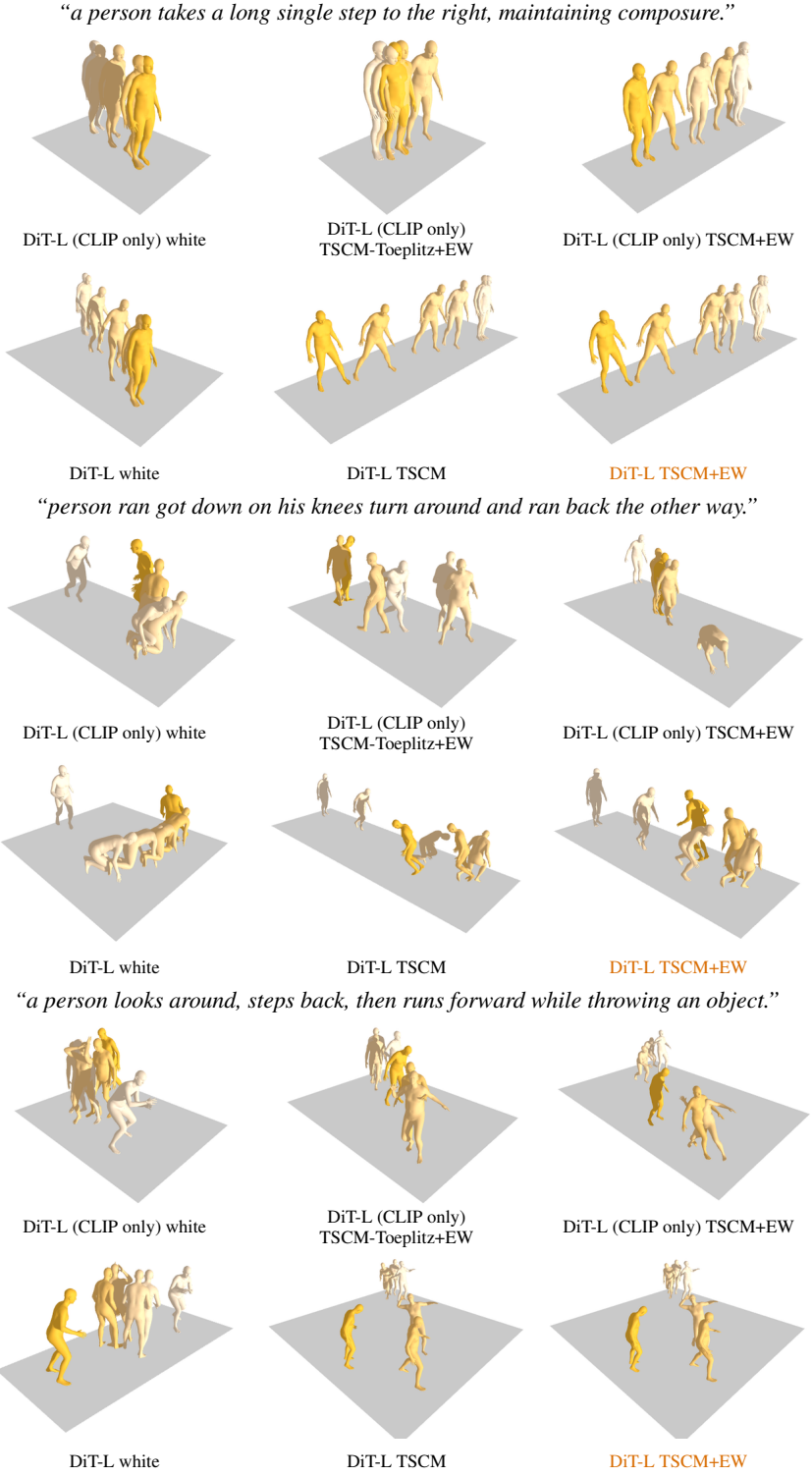

Figure 13: **Text-to-motion variants for three shared captions.** Each caption has a CLIP-only DiT-L row and a CLIP+T5 row; the latter is called DiT-L in Table 3. EW denotes early-window guidance. Within the T5 row, white versus TSCM uses the same conditioning recipe, and TSCM versus TSCM+EW changes the sampling schedule. The CLIP-only row changes both source and guidance relative to white noise. Settings and quantitative comparisons: Appendix B.1, Table 3, and Table 7.

Music-to-dance: released and fine-tuned FlowerDance.

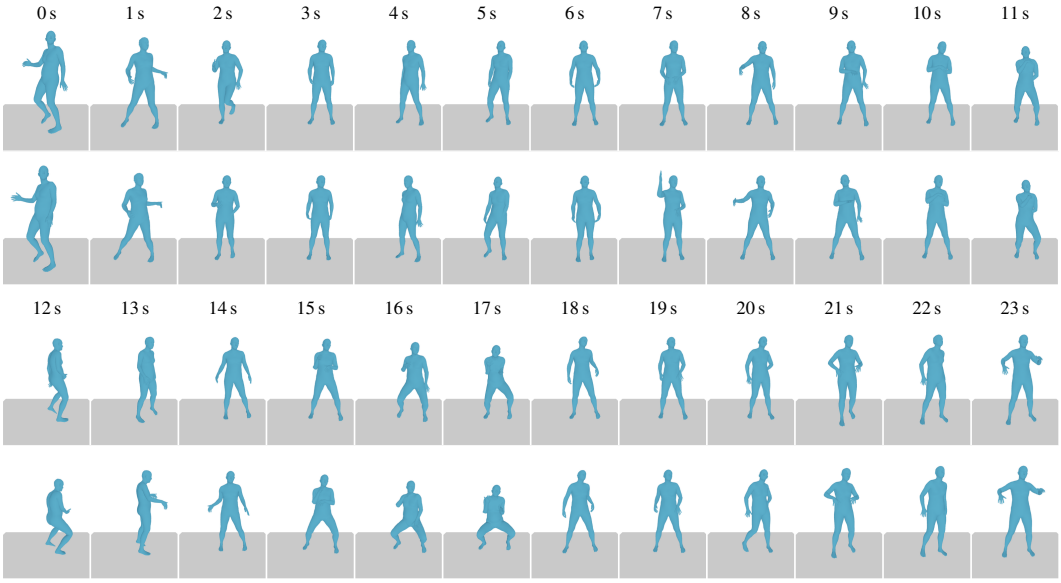

Figure 14: **FlowerDance on FineDance test slice `211_slice3`.** In each time block, the top row is the released FlowerDance checkpoint and the bottom row is its 10-epoch fine-tune with a TSCM source. Columns show front-view poses at one-second intervals for the same music slice, retaining all 24 time points from 0 to 23 s. The two rows differ in both training budget and source; this illustration does not isolate the source effect. Table 17 provides the matched white-source fine-tune.

Text-to-video: matched source replacement.

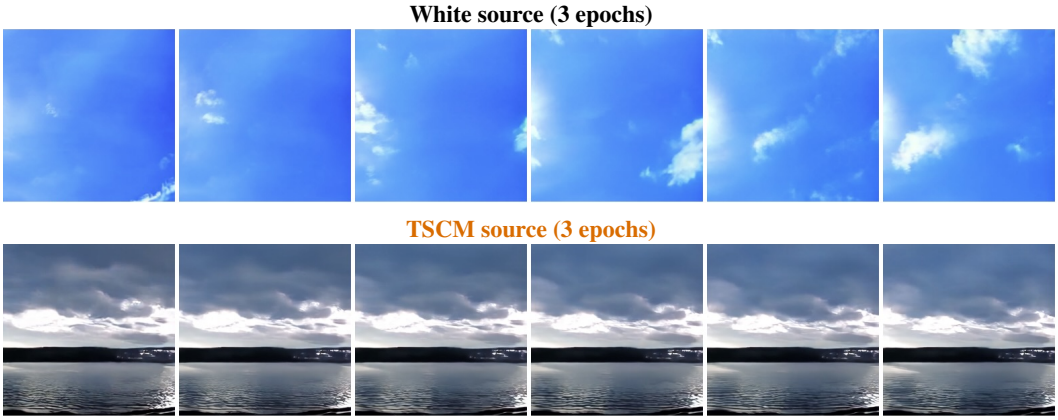

Figure 15: **Wan2.1 sky time-lapse examples after three fine-tuning epochs.** Top: white source; bottom: TSCM source. The model and fine-tuning recipe are shared, as in Appendix B.17. Both use the caption "A time-lapse video of the sky with clouds drifting and changing shape." Each row presents six frames from one generated clip in temporal order, showing the evolution of the cloud pattern within that sample. The two outputs depict different sky scenes under this shared caption; Table 2 evaluates the generated distributions with FVD.

C Related work

Temporal and spectral source structure. TSFlow (Kollovieh et al., 2025) is a direct temporal-source predecessor: its published constructions use squared-exponential, Ornstein–Uhlenbeck, and periodic Gaussian-process kernels, including conditional GP priors for forecasting. SpecGrad (Koizumi et al., 2022) shapes diffusion noise with a time-varying filter derived from the conditioning spectrum, and PYoCo (Ge et al., 2023) uses shared or progressive noise to correlate video frames. TimeBridge (Park et al., 2026) studies empirical per-position diagonal moments and a GP extension. TSCM differs from these concrete constructions by estimating a complete temporal matrix for every channel and supported length, rather than fixing a kernel or a prescribed correlation pattern. It can retain nonstationary within-window statistics and level–fluctuation cross covariance. This is a difference in the fitted source, not a claim that general GP or flow-matching frameworks cannot represent such a matrix.

Condition-adaptive Gaussian sources. PriorGrad (Lee et al., 2022) uses mel-energy-derived diagonal variances for waveform synthesis and phoneme-indexed means and variances for acoustic modeling. Its vocoder prior is zero-mean, so zero mean is not itself a distinction of TSCM. CW-Gen (Yang et al., 2026) learns a Joint Mean–Covariance Estimator (JMCE) from historical context, producing a mean and a sliding-window cross-variable covariance for each future position. CW-Diff operates in conditionally whitened coordinates, whereas CW-Flow directly interpolates between data and the condition-adaptive Gaussian source. TSCM likewise operates in the generator's data coordinates, but uses offline length-indexed statistics rather than a condition-dependent moment predictor.

Covariance axes. Let i, j denote sequence positions and c, d denote channels. The stated CW-Gen construction applies a separate cross-variable factor to independent Gaussian innovations at each position, giving

$$\mathrm{Cov}(x_{0,i,c}, x_{0,j,d} \mid y) = \mathbf{1}_{\{i=j\}}[\widehat{\Sigma}_i(y)]_{cd}. \tag{C.1}$$

TSCM instead gives

$$\mathrm{Cov}(x_{0,i,c}, x_{0,j,d} \mid L, y) = \mathbf{1}_{\{c=d\}}[A_{c,L}]_{ij}. \tag{C.2}$$

These sources retain different covariance blocks: within-position cross-variable dependence versus within-channel cross-position dependence. The distinction concerns the source, not the dependencies representable by the learned generator. Neither block structure is the unrestricted $LC \times LC$ covariance, and their different axes do not by themselves establish that arbitrary compositions commute.

Finite-step integration versus training and prior proximity. PriorGrad's analysis includes an ELBO comparison and optimization-Hessian conditioning in a linear model (Lee et al., 2022). CW-Gen gives sufficient conditions for estimated conditional moments to reduce data-to-prior KL divergence relative to a standard Gaussian (Yang et al., 2026). Preconditioned Flow Matching (Ahamed et al., 2026) derives the Gaussian population field and studies covariance-induced anisotropy in velocity-regression optimization, using preconditioning to improve that geometry. These are distinct from our numerical objective: even with the covariance-driven population field available exactly, a finite-step integrator incurs source-dependent error. Our analysis diagonalizes that field for arbitrary positive-definite source and target covariances in their generalized eigenbasis, derives its exact uniform-Euler modal errors, and minimizes their aggregate risk. The resulting optimum holds for every $N \geq 2$, not only asymptotically. A local stability result quantifies second-order excess risk near the match. It addresses the numerical cost of approximate covariance matching, rather than asserting improved KL divergence or faster training.

Empirical estimation and structural approximation. TSCM uses Ledoit–Wolf shrinkage (Ledoit & Wolf, 2004) and Cholesky sampling, both standard components. Its source contribution is the length-indexed estimation and use of full within-channel temporal statistics under the finite-step criterion. At each length, one window per eligible training clip avoids weighting a clip in proportion to its number of available windows. The estimator does not subtract each window's temporal mean, so level variation and its interaction with fluctuations remain represented. By contrast, TSCM-Toeplitz uses a centered-ACF fluctuation block and a separately estimated level with zero cross covariance; it is a structured approximation evaluated against the same numerical target.

Residual population means, estimation error, and shrinkage must therefore be distinguished from the structural restrictions imposed by a source family.

Paths, couplings, and numerical schemes. Flow matching, rectified flow, and stochastic interpolants study the construction of generative paths (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023); minibatch optimal transport changes the endpoint pairing (Tong et al., 2024). Our primary source optimization fixes independent pairing and a linear interpolation; the correlated extensions retain an explicit aligned-coupling assumption. The scope matters: for a scalar Gaussian optimal-transport coupling, trajectories are affine and Euler is exact regardless of the variance ratio. The exact matched scale is specific to uniform Euler. Our extensions separate target covariance shape from solver-dependent scalar calibration and establish leading endpoint-error cancellation for even-order Runge–Kutta methods on symmetric paths and grids, including an aligned correlated-coupling family. The marginal covariance need not optimize each conditional field; our average conditional regret bound quantifies the cost through condition-mean variation and covariance heterogeneity. A separate local minimax theorem covers non-Gaussian targets without claiming a non-Gaussian endpoint optimum. These boundaries separate the numerical result from universal claims about generative-model quality.

Guidance schedules and initial covariance bias. Limiting guidance to an interval improves diffusion sampling (Kynkäänniemi et al., 2024), and stage-wise CFG analysis relates early mean-directed dynamics to later diversity loss (Jin et al., 2026). The predictor–corrector interpretation provides another account of guided dynamics (Bradley & Nakkiran, 2024). At a fixed length, our contribution is an initial-time identity under independent coupling, $\Delta(x, 0, y) = \mu_y - \mu$, and its exact-integration consequence

$$\mathrm{Cov}(X^{(w)}(t)) - \Sigma_t = t^2 w(w-1)\,\mathrm{Cov}_Y(\mu_Y) + o(t^2), \tag{C.3}$$

under the stated regularity assumptions and constant guidance near the start. This identifies the between-condition component of the local covariance bias and connects its level projection to temporal source calibration. It motivates an early-window schedule, but does not assert that the excess necessarily persists to the endpoint or that weakening early guidance is universally optimal. The measured duration and effects on covariance and distributional metrics are empirical parts of the contribution.

Evaluation backbones and representations. MDM (Tevet et al., 2023), MotionDiffuse (Zhang et al., 2024), MLD (Chen et al., 2023), MoMask (Guo et al., 2024) and Light-T2M (Zeng et al., 2025) provide architectural context for our text-conditioned sequence testbed. Our matched MDM and Light-T2M reruns evaluate source replacement within existing generators. MoRAE (Zhu et al., 2026) diagnoses the within-frame ill-conditioning of motion latents and builds a compact, near-isotropic latent for flow matching; that work conditions the channel axis, this one conditions time, and Section B.16 shows the two compose. Our contribution is orthogonal to the generator architectures: it changes the noise they start from and the guidance they apply in the first steps.