ENTROKINE3D: SHARED KINETIC-PROFILE SUPERVISION FOR SINGLE-REFERENCE 3D DANCE CAMERA SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic 3D dance camera trajectory synthesis should support diverse artistic interpretations of the same choreography and accompanying music through variations in viewpoint, framing, and visual rhythm. However, currently available training data only provide one authored camera trajectory for each exact performance, and this observation is one accepted realization rather than the unique valid answer, making it difficult for instance-level supervision to characterize plausible variation across camera trajectories while preserving high-fidelity motion. In this paper, we propose **EntroKine3D**, a novel framework that supplements single-reference supervision with shared kinetic-profile constraints pooled from related performance states for high-quality stochastic synthesis. Specifically, to address the single-reference supervision challenge, we design an Entropy-calibrated Kinetic Profile Modeling (E-KPM) module to supervise generated camera windows against structured population-level kinetic references built from single-reference camera-motion statistics. E-KPM aligns continuous kinetic distributions via Sinkhorn divergence and matches discrete motion-state occupancy, condition-dependent allocation, and temporal transitions. To balance trajectory diversity with kinetic high-fidelity, we extend entropy calibration to further match profile uncertainty and the information that the past performance state provides about future camera-motion states to reference statistics under finite-sample, finite-step optimization. Furthermore, we incorporate latent consistency distillation as an efficient instantiation for quick four-step generation. On the DCM dataset, EntroKine3D achieves FID-K/FID-S of $4.342^{\pm 0.22}/0.398^{\pm 0.03}$ and DIV-K of $3.072^{\pm 0.12}$, close to the held-out reference value 3.290. Same-condition evaluations show that entropy calibration yields diverse yet faithful camera trajectories under one performance, while four-step generation achieves a $9.39\times$ model-level speedup over the 50-step teacher.

1 INTRODUCTION

Camera movement shapes the visual presentation of a 3D dance performance by controlling viewpoint, framing, and visual rhythm (Wang et al., 2024a;b; Xu et al., 2025). The same choreography and music admit multiple reasonable cinematographic interpretations: different camera trajectories can emphasize different dancers, spatial relations, or rhythmic accents while remaining equally plausible. The goal of automatic camera synthesis is therefore not merely to produce different outputs, but to offer *distinct yet plausible alternatives* for a fixed performance.

Learning-based camera synthesis is trained on datasets such as DCM (Wang et al., 2024a), which provide exactly one authored trajectory for each exact music–dance performance. This observed trajectory is one accepted realization, not the unique correct answer. Instance-level trajectory supervision rewards matching the observed realization, but provides no direct supervision for unobserved yet plausible alternatives: nothing in the training signal specifies which deviations from the observation remain valid camerawork. Stochastic sampling alone does not close this gap: injected noise produces differences without separating meaningful variation from unstructured deviation.

Our key observation is that kinetic patterns are far more reusable across performances than exact trajectories. Descriptions of how the camera moves—movement intensity, state occupancy, 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.

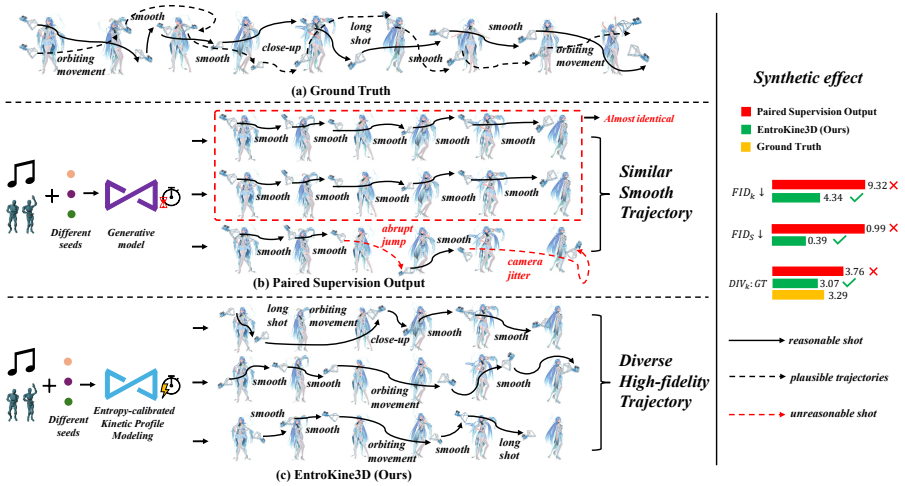


Figure 1: **From single trajectories to shared kinetic profiles.** (a) For each exact music–dance performance, the dataset provides one authored camera trajectory (solid arrows) among many plausible alternative trajectories (dashed arrows). (b) A generator trained with paired supervision produces almost identical, uniformly smooth trajectories across sampling seeds, with occasional unreasonable shots (abrupt jump, camera jitter; red). (c) EntroKine3D produces diverse yet high-fidelity trajectories for the same input, varying shot usage (long shot, orbiting movement, close-up) while keeping shots reasonable throughout. The right panel summarizes the synthetic effect: EntroKine3D attains lower FID-K/FID-S and DIV-K closer to ground truth.

transition patterns—recur across related performance states, even though absolute viewpoints and shot locations do not. Kinetic statistics are largely insensitive to where a shot takes place while remaining sensitive to how the camera moves, so they transfer across related performances far better than raw trajectories. This suggests supplementing single-reference supervision with *shared kinetic references*: population-level statistics pooled from training cameras at similar performance states (Figure 1).

We instantiate this supervision in **EntroKine3D**, whose core component is Entropy-Calibrated Kinetic Profile Modeling (E-KPM). E-KPM builds a structured kinetic profile of camera behavior and matches it between generated and reference populations at four levels: continuous kinetic geometry, which preserves the support of camera behaviors (Feydy et al., 2019); discrete state occupancy, which describes *what* camera behaviors occur; conditional occupancy, which describes *under which performance state* they occur; and temporal transition structure, which describes *how* they evolve. On top of profile matching, explicit calibration directly optimizes selected profile-uncertainty statistics toward their reference values (Shannon, 1948; Cover & Thomas, 2006; Schreiber, 2000), providing additional, non-collinear optimization signals under finite-sample, finite-step training. The supervision is deployed on a latent consistency generator (Song et al., 2023; Luo et al., 2024), so each alternative trajectory costs only four network evaluations. Experiments on DCM show that EntroKine3D improves distributional quality over instance-level and set-level supervision controls (FID-K 4.342 versus 6.381 for paired supervision), broadens the fixed-condition stochastic response (NS 0.34 versus 0.08), and is preferred in same-condition human set evaluations, while four-step sampling gives a $9.39\times$ speedup over the diffusion teacher. Our contributions are:

- We cast single-reference dance camera synthesis as a distributional supervision problem: the observed camera is one accepted realization; plausible alternatives stay unobserved. We show that instance-level supervision leaves the fixed-condition stochastic response narrow and poorly supported by reference camerawork.
- We introduce shared structured kinetic-profile supervision (E-KPM), which pools reusable camera-motion statistics across related performance states and matches continuous kinetic support, condition-dependent allocation, and temporal evolution between generated and reference profiles.

- We explicitly calibrate profile uncertainty during finite optimization and validate the resulting quality–variation balance through matched controls, temporal interventions, fixed-condition sampling, and human set evaluation. The supervision runs on a four-step latent generator for efficient multi-sample synthesis.

2 Related Work

2.1 Camera and Music-Conditioned Motion Synthesis

Early virtual cinematography used geometric constraints or learned camera behaviors from filmed examples (Christie et al., 2008; Lino & Christie, 2015; Jiang et al., 2020); more recent work studies text-, scene-, and character-conditioned camera generation (Jiang et al., 2024; Courant et al., 2024; Li et al., 2024b; Zhang et al., 2025b; Courant et al., 2026; Dehghanian et al., 2025). Music-driven motion synthesis has likewise progressed from autoregressive and discrete models to diffusion and masked generation for detailed or long-form dance (Huang et al., 2021; Li et al., 2021; Siyao et al., 2022; Tevet et al., 2023; Tseng et al., 2023; Guo et al., 2024; Li et al., 2023; 2024a; Zhang et al., 2025a). For dance cameras, DanceCamera3D, DanceCamAnimator, and TemMEGA improve generation quality, control, and temporal modeling (Wang et al., 2024a;b; Xu et al., 2025). DanceStyleCam (Huang et al., 2026) instead studies explicit style-conditioned camera synthesis. EntroKine3D targets stochastic alternatives without an explicit style signal, focusing on how to supervise plausible camera variation when each exact performance provides only one authored trajectory.

2.2 Distributional Supervision and Efficient Synthesis

Distributional objectives constrain generated samples beyond per-example targets (Tolstikhin et al., 2018; Genevay et al., 2018; 2019; Feydy et al., 2019). E-KPM combines continuous alignment with a fixed soft kinetic codebook (Lloyd, 1982; Arthur & Vassilvitskii, 2007; van den Oord et al., 2017) and explicitly calibrates uncertainty statistics of the resulting profiles (Shannon, 1948; Cover & Thomas, 2006). We use diffusion and consistency distillation (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Rombach et al., 2022; Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2024; Song & Dhariwal, 2024) only as an efficient sampling backbone.

3 Method

Given synchronized music features $A = \{\mathbf{a}_t\}_{t=1}^{T}$ and 3D dance motion $M = \{\mathbf{m}_t\}_{t=1}^{T}$, our goal is to synthesize a camera trajectory $C = \{\mathbf{c}_t\}_{t=1}^{T}$, where $\mathbf{c}_t = (\mathbf{e}_t, \mathbf{R}_t, f_t)$ consists of the camera eye position $\mathbf{e}_t \in \mathbb{R}^3$, orientation $\mathbf{R}_t \in SO(3)$, and field of view f_t. With a stochastic sampling stream ϵ, the synthesis process is written as

$$\widehat{C} = G_\theta(A, M, \epsilon). \tag{1}$$

Figure 2 illustrates the overall framework of **EntroKine3D**. Training cameras are first summarized into shared kinetic profiles over related music–dance performance states. A conditional latent camera generator then produces stochastic camera trajectories from the synchronized music and dance. E-KPM supervises the decoded predictions by matching their kinetic support, condition-dependent state allocation, temporal organization, and profile uncertainty to frozen train-only references. The trained student is finally deployed with four network evaluations for efficient multi-sample synthesis.

3.1 Shared Kinetic Profile Construction

Training data contain a single authored camera for each exact music–dance performance. Instead of constructing additional trajectory targets, we summarize reusable camera-motion regularities from the training set and use them as shared kinetic references. The construction contains a continuous kinetic representation, an input-only performance partition, and a frozen kinetic-state codebook.

Camera kinetic representation. Absolute camera trajectories depend strongly on viewpoint and spatial configuration. We therefore characterize each camera window by how its degrees of freedom

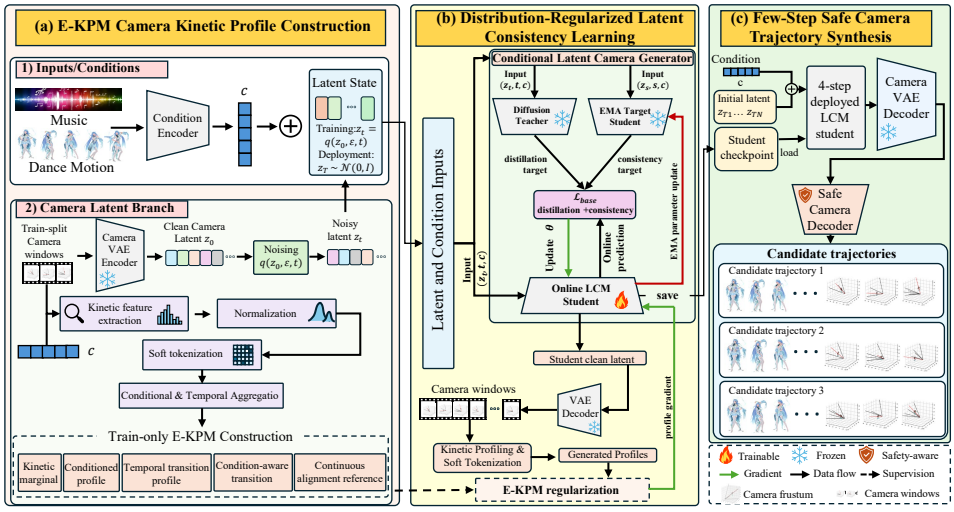


Figure 2: **EntroKine3D framework.** (a) Training cameras are converted into continuous kinetic descriptors, performance states, and soft kinetic states to construct frozen shared kinetic profiles. (b) A music–dance-conditioned latent consistency student predicts a clean camera latent, which is decoded into camera space. (c) E-KPM compares the generated and reference kinetic profiles through continuous support, context-dependent state allocation, temporal organization, and uncertainty calibration. (d) At inference, different stochastic streams are propagated through the four-step student to synthesize multiple camera trajectories for the same music–dance performance.

move rather than by their absolute values. For the i-th camera window, we define the kinetic descriptor

$$\begin{aligned}\mathbf{K}_i = [&V_i(\mathbf{e}), A_i(\mathbf{e}), V_i(\mathbf{b}^z), A_i(\mathbf{b}^z), V_i(\mathbf{b}^y), A_i(\mathbf{b}^y),\\ &V_i(\mathbf{b}^x), A_i(\mathbf{b}^x), V_i(f), A_i(f)] \in \mathbb{R}^{10},\end{aligned} \tag{2}$$

where $\mathbf{R}_t = [\mathbf{b}_t^x, \mathbf{b}_t^y, \mathbf{b}_t^z]$, and $V_i(\cdot)$ and $A_i(\cdot)$ measure window-level velocity and acceleration magnitude, respectively. Thus, $\mathbf{K}_i$ summarizes camera dynamics while reducing sensitivity to its absolute shot location. The exact derivative operators and normalization are given in Appendix A.

Performance-state partition. Shared statistics should be pooled among music–dance contexts with similar dynamics. We construct an input-only descriptor $\Phi_Y(A, M)_i$ from audio energy, onset activity, and normalized dancer-motion statistics, and assign every window to one of N_Y frozen performance states:

$$Y_i = Q_Y(A, M)_i = \arg\min_{y \in \{1,\dots,N_Y\}} \left\| \widetilde{\Phi}_Y(A, M)_i - \boldsymbol{\nu}_y \right\|_2^2. \tag{3}$$

The centroids $\{\boldsymbol{\nu}_y\}$ are fitted on training data only. Because Y_i is computed exclusively from music and dance, it organizes camera references by performance context without reading camera outputs. The deployed generator still receives the full continuous music–dance condition; Y_i is used only for kinetic-profile construction and supervision.

Soft kinetic states. To complement the continuous descriptor, we fit a frozen codebook of N_κ kinetic centers $\{\boldsymbol{\mu}_k\}_{k=1}^{N_\kappa}$. A standardized descriptor $\widetilde{\mathbf{K}}_i$ is softly assigned to the codebook as

$$q_i(k) = \frac{\exp\left(-\|\widetilde{\mathbf{K}}_i - \boldsymbol{\mu}_k\|_2^2/\tau\right)}{\sum_{j=1}^{N_\kappa} \exp\left(-\|\widetilde{\mathbf{K}}_i - \boldsymbol{\mu}_j\|_2^2/\tau\right)}. \tag{4}$$

The continuous descriptor preserves kinetic geometry, while the soft states provide a compact representation for occupancy and temporal organization.

Frozen reference profiles. Using the training-camera windows, we estimate four types of frozen reference statistics: the continuous kinetic distribution $P_r(\mathbf{K})$, the global state occupancy $p_r(k)$, the performance-conditioned occupancy $p_r(k \mid Y)$, and the temporal joint $J_{r,\ell}(Y^-, K^-, K^+)$. They respectively describe continuous kinetic support, overall state usage, context-dependent state allocation, and ordered camera-state evolution. All normalizers, performance-state centroids, kinetic centers, supports, and reference statistics are fitted once on the training split and remain fixed afterward. Detailed estimation is given in Appendix A.

3.2 Conditional Latent Camera Generator

EntroKine3D realizes stochastic camera synthesis in a learned latent space. A condition encoder first converts the complete synchronized music–dance sequence into $\mathbf{c} = E_{\text{cond}}(A, M)$. A frozen camera VAE encodes an authored trajectory C into the clean latent $\mathbf{z}_0 = E_{\text{vae}}(C)$. Following the standard diffusion construction, a noisy latent at timestep t is

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{5}$$

We use a frozen conditional diffusion teacher F_T and distill it into an online latent-consistency student F_θ. Given $(\mathbf{z}_t, t, \mathbf{c})$, the student predicts a clean camera latent, which is decoded into camera space:

$$\widehat{\mathbf{z}}_{0,t} = F_\theta(\mathbf{z}_t, t, \mathbf{c}), \qquad \widehat{C}_t = D_{\text{vae}}(\widehat{\mathbf{z}}_{0,t})\,. \tag{6}$$

The decoded prediction $\widehat{C}_t$ is the camera trajectory to which E-KPM is applied.

The student receives two latent-space targets. The frozen teacher provides a clean-latent distillation target, while an exponential-moving-average student provides an adjacent-noise-level consistency target. Their weighted objective is $\mathcal{L}_{\text{base}} = \lambda_{x_0}\mathcal{L}_{x_0} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}$. The noising schedule, teacher transport, EMA update, and exact definitions of $\mathcal{L}_{x_0}$ and $\mathcal{L}_{\text{cons}}$ are given in Appendix B.

3.3 Entropy-Calibrated Kinetic Profile Supervision

E-KPM connects the frozen training profiles defined in Section 3.1 to the stochastic camera generator. For every decoded student prediction $\widehat{C}$, we apply the same frozen kinetic extractor, performance-state partition, and codebook as for the reference cameras. This produces generated continuous descriptors, soft kinetic states, occupancy profiles, and temporally ordered state statistics. We compare the generated and reference profiles in terms of continuous kinetic support, context-dependent state allocation, and temporal organization.

Continuous kinetic support. Let $P_g(\mathbf{K})$ and $P_r(\mathbf{K})$ denote the generated and reference distributions of standardized continuous kinetic descriptors. We align them with the debiased Sinkhorn discrepancy

$$\mathcal{D}_{\text{SA}} = \text{OT}_\varepsilon(P_g, P_r) - \frac{1}{2}\,\text{OT}_\varepsilon(P_g, P_g) - \frac{1}{2}\,\text{OT}_\varepsilon(P_r, P_r). \tag{7}$$

This geometry-aware term preserves relationships in the continuous kinetic space that discrete state counts alone discard.

Contextual state allocation. Soft assignments from Eq. equation 4 define the generated and reference state distributions $p_d(k)$ and $p_d(k \mid Y)$ for $d \in \{r, g\}$. We match global state usage and its allocation across performance contexts by

$$\mathcal{D}_M = D_{\text{KL}}(p_r(k) \| p_g(k))\,, \tag{8}$$

$$\mathcal{D}_C = \sum_y \omega_y^{\text{occ}} D_{\text{KL}}(p_r(k \mid y) \| p_g(k \mid y))\,. \tag{9}$$

Here ω_y^{occ} is the reference-state mixture fitted on the training split.

Temporal organization. Static occupancy does not specify the order in which camera states evolve. For each eligible source window, we pair a past music–dance state Y_i^- with the current kinetic state $K_{d,i}^-$ and a future kinetic state $K_{d,i,\ell}^+$ at lag $\ell \in \Lambda_{\text{TE}}$. Their soft assignments define the joint count

$$N_{d,\ell}(y, k, k') = \sum_{i \in \mathcal{A}_{d,\ell}} \mathbb{1}[Y_i^- = y]\, q_{d,i}^-(k)\, q_{d,i,\ell}^+(k'), \tag{10}$$

which is normalized into a joint distribution $J_{d,\ell}(y, k, k')$. From this joint we obtain the state-transition distribution $T_{d,\ell}(k' \mid k)$ and the performance-conditioned transition distribution $T^{Y^-}_{d,\ell}(k' \mid k, y)$. We match them through

$$\mathcal{D}_T = \sum_{\ell \in \Lambda_{\mathrm{TE}}} \omega_\ell \sum_k \pi_{r,\ell}(k) D_{\mathrm{KL}}(T_{r,\ell}(\cdot \mid k) \| T_{g,\ell}(\cdot \mid k)), \tag{11}$$

$$\mathcal{D}_{CT} = \sum_{\ell \in \Lambda_{\mathrm{TE}}} \omega_\ell \sum_{y,k} \pi_{r,\ell}(y, k) D_{\mathrm{KL}}\Big(T^{Y^-}_{r,\ell}(\cdot \mid k, y) \| T^{Y^-}_{g,\ell}(\cdot \mid k, y)\Big). \tag{12}$$

Thus, $\mathcal{D}_T$ preserves how camera states evolve, while $\mathcal{D}_{CT}$ preserves how this evolution depends on the preceding music–dance context.

Combining continuous support, state allocation, and temporal organization gives the structured kinetic-profile objective

$$\mathcal{L}_{\mathrm{profile}} = \lambda_{\mathrm{SA}} \mathcal{D}_{\mathrm{SA}} + \lambda_M \mathcal{D}_M + \lambda_C \mathcal{D}_C + \lambda_T \mathcal{D}_T + \lambda_{CT} \mathcal{D}_{CT}. \tag{13}$$

Uncertainty calibration. In addition to matching profile distributions, we explicitly calibrate the amount and condition dependence of kinetic uncertainty. For the temporal joint at lag ℓ,

$$H^d_{\mathrm{tr},\ell} = H_{J_{d,\ell}}(K^+ \mid K^-), \qquad H^d_{\mathrm{cond},\ell} = H_{J_{d,\ell}}(K^+ \mid K^-, Y^-), \tag{14}$$

$$\mathrm{TE}_{d,\ell} = I_{J_{d,\ell}}(Y^-; K^+ \mid K^-) = H^d_{\mathrm{tr},\ell} - H^d_{\mathrm{cond},\ell}. \tag{15}$$

The first two quantities measure transition uncertainty with and without the past performance state, while $\mathrm{TE}_{d,\ell}$ measures how strongly the past condition reduces uncertainty about future camera motion.

We match the generated marginal, conditional, and temporal uncertainty statistics to their frozen reference values:

$$\mathcal{L}_{\mathrm{entropy}} = \gamma_M \Delta H_M + \gamma_C \Delta H_C + \gamma_{\mathrm{tr}} \Delta H_{\mathrm{tr}} + \gamma_{CT} \Delta H_{\mathrm{cond}} + \gamma_{\mathrm{TE}} \Delta_{\mathrm{TE}}, \tag{16}$$

$$\Delta_{\mathrm{TE}} = \sum_{\ell \in \Lambda_{\mathrm{TE}}} \omega_\ell \left| \mathrm{TE}_{r,\ell} - \mathrm{TE}_{g,\ell} \right|. \tag{17}$$

The complete E-KPM supervision is $\mathcal{L}_{\text{E-KPM}} = \mathcal{L}_{\mathrm{profile}} + \mathcal{L}_{\mathrm{entropy}}$. Detailed entropy estimators, temporal supports, smoothing rules, and the factorization of the temporal joint are provided in Appendix A.

3.4 Joint Training and Four-Step Synthesis

The online student is jointly optimized with $\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \mathcal{L}_{\text{E-KPM}}$. During each update, the online student receives the noisy camera latent and the complete music–dance condition. Its clean-latent prediction is decoded by the frozen VAE and evaluated by E-KPM against the frozen train-only kinetic references. Gradients from E-KPM pass through the frozen decoder to the online student. The teacher, EMA target, camera VAE, performance-state partition, kinetic codebook, and reference profiles remain fixed.

Four-step synthesis. At inference, the condition encoder first computes $\mathbf{c} = E_{\mathrm{cond}}(A, M)$. A sampling stream starts from Gaussian latent noise and is propagated through the student using four network evaluations at $(999, 666, 333, 0)$. The variance-corrected update between adjacent noise levels is given in Appendix B. The final clean latent is decoded to camera parameters and passed through the deterministic geometry-safe projection, which orthonormalizes orientation and repairs invalid geometric views.

Different sampling streams vary the initial latent and intermediate Gaussian perturbations while keeping the music–dance condition fixed. Thus, one trained model directly produces $\{\widehat{C}^{(r)} = G_\theta(A, M, \epsilon_r)\}_{r=1}^R$ for the same performance without retraining or additional control signals.

4 Experiments

We evaluate EntroKine3D on DCM and analyze supervision formulation, temporal profile modeling, uncertainty calibration, same-condition alternatives, and few-step sampling.

Table 1: Grouped DCM test comparison under the common camera representation and geometry-safe projection. Superscripts are standard deviations over three checkpoint means, each averaged over three synthesis streams; point rows use one available checkpoint. DIV is compared with GT. Bold marks the best reported mean, with diversity ranked by distance to GT.

Method	Quality		Diversity		Dancer fidelity		Kinetic / temporal diagnostics		
	$FID_k \downarrow$	$FID_s \downarrow$	$DIV_k \approx GT$	$DIV_s \approx GT$	DMR ↓	LCD ↓	SK ↓	Coverage ↑	CJ ↓
GT	–	–	3.290	1.74	–	–	–	–	–
DanceCamera3D (Wang et al., 2024a)	9.555	0.890	3.848	1.16	0.0044	0.245	5.52	0.59	1.067
DanceCamAnimator (Wang et al., 2024b)	9.320	0.993	3.761	2.22	0.0037	**0.189**	4.74	0.75	0.976
LODGE-Cam (Li et al., 2024a)	17.420	1.782	2.280	1.48	0.0040	0.218	15.67	0.25	1.584
EDMG-Cam (Zhang et al., 2025a)	14.860	1.536	**3.105**	**1.66**	0.0043	0.226	12.83	0.42	1.296
EDGE-Cam (Tseng et al., 2023)	13.284	1.394	3.566	1.94	0.0045	0.231	9.91	0.57	1.391
Internal backbone and mechanism controls									
Diffusion Teacher	7.546	0.603	2.697	1.32	0.0053	0.264	8.01	0.53	0.581
LCM Student w/o E-KPM	$7.711^{\pm 0.48}$	$0.632^{\pm 0.05}$	$2.769^{\pm 0.15}$	$1.54^{\pm 0.11}$	$0.0045^{\pm 6\times 10^{-4}}$	$0.219^{\pm 0.01}$	$6.46^{\pm 0.44}$	$0.52^{\pm 0.04}$	$0.542^{\pm 0.06}$
EntroKine3D w/o entropy calibration	$5.072^{\pm 0.12}$	$0.410^{\pm 0.09}$	$2.914^{\pm 0.08}$	$1.59^{\pm 0.14}$	$0.0037^{\pm 6\times 10^{-4}}$	$0.211^{\pm 0.014}$	$\mathbf{2.57^{\pm 0.25}}$	$0.84^{\pm 0.07}$	$\mathbf{0.118^{\pm 0.03}}$
EntroKine3D	$\mathbf{4.342^{\pm 0.22}}$	$\mathbf{0.398^{\pm 0.03}}$	$3.072^{\pm 0.12}$	$1.62^{\pm 0.08}$	$\mathbf{0.0032^{\pm 4\times 10^{-4}}}$	$0.208^{\pm 0.01}$	$2.65^{\pm 0.21}$	$\mathbf{0.89^{\pm 0.03}}$	$0.122^{\pm 0.02}$

4.1 Experimental Setup

Dataset. DCM provides synchronized music, 3D dance, and authored cameras (Wang et al., 2024a). The grouped protocol splits raw sequences into 81/41/49 train/validation/test sequences before windowing; the test split lasts about ten minutes. All normalizers, condition states, codebooks, and targets use training data. Appendix D.6 gives the short-clip protocol.

Evaluation metrics. FID-K/FID-S and DIV-K/DIV-S measure distributional quality and diversity relative to GT, while DMR/LCD measure dancer fidelity (Wang et al., 2024a;b; Heusel et al., 2017). DIV is judged by agreement with GT. SK, Coverage, CJ, and profile discrepancies diagnose kinetic structure; NS measures fixed-condition stochastic response. Appendices C and E define metrics. E-KPM directly regularizes camera kinetics; cinematographic properties (framing, visibility, shot continuity, and event timing) are evaluated independently in Appendix E, not implied by kinetic agreement.

Implementation details. Validation selects $W_{\text{prof}} = 75$, $\eta = 0.5$, and $a_\gamma = 1$. Training uses four RTX PRO 5000 GPUs for 3,000 epochs with batch size 64; temporal profiles use 75-frame anchors and lags $\{75, 150\}$. All methods share the same representation, renderer, and safe projection. Repeated rows average three streams over three checkpoints; uncertainty conventions and complete protocols are given in Appendix G.

4.2 Comparison with State-of-the-Art Methods

Baselines include DanceCamera3D and DanceCamAnimator (Wang et al., 2024a;b); EDGE-Cam, LODGE-Cam, and EDMG-Cam adapt motion-generation backbones to the shared camera representation (Tseng et al., 2023; Li et al., 2024a; Zhang et al., 2025a). TemMEGA appears under the original short-clip protocol in Appendix D.6.

As shown in Table 1, EntroKine3D achieves FID-K $4.342^{\pm 0.22}$ and FID-S $0.398^{\pm 0.03}$, with DIV-K/DIV-S close to GT. DanceCamAnimator retains the lowest LCD, and EDMG-Cam has closer diversity. Relative to LCM-only, E-KPM reduces FID-K and CJ while increasing Coverage. Independent cinematography diagnostics also favor Full over entropy-free (Appendix E, Table 26).

4.3 Analysis of EntroKine3D

Supervision formulation. T0 uses paired sample-wise supervision; T1/T2 draw non-paired references globally or from the same profiling state; T3 uses state-conditioned set matching. All variants share the four-step backbone. T0 retains strong point-wise dancer fidelity but has substantially narrower diversity and lower reference-state coverage, while T3 gives a better quality–variation balance

Table 2: Controlled supervision and E-KPM ablations. Superscripts are standard deviations over three checkpoint means ($a^{\pm b}$), as in Table 1.

Variant	FID-K ↓	FID-S ↓	DIV-K ≈GT	SK ↓	Coverage ↑	CJ ↓	C.KL ↓	TE.g ↓
Supervision formulation (same four-step backbone)								
T0: Paired	$6.381^{\pm 0.13}$	$0.429^{\pm 0.06}$	$2.032^{\pm 0.19}$	$4.73^{\pm 0.26}$	$0.49^{\pm 0.025}$	$0.267^{\pm 0.021}$	$2.65^{\pm 0.063}$	$0.73^{\pm 0.06}$
T1: Global set	$6.408^{\pm 0.38}$	$0.735^{\pm 0.09}$	$2.729^{\pm 0.16}$	$5.24^{\pm 0.45}$	$0.61^{\pm 0.033}$	$0.712^{\pm 0.058}$	$2.90^{\pm 0.063}$	$0.66^{\pm 0.07}$
T2: Condition-bin set	$5.486^{\pm 0.25}$	$0.595^{\pm 0.08}$	$2.887^{\pm 0.14}$	$3.75^{\pm 0.27}$	$0.73^{\pm 0.036}$	$0.481^{\pm 0.042}$	$\mathbf{0.42^{\pm 0.044}}$	$0.42^{\pm 0.05}$
T3: Set distribution	$4.851^{\pm 0.29}$	$0.477^{\pm 0.05}$	$2.995^{\pm 0.12}$	$3.15^{\pm 0.21}$	$0.81^{\pm 0.031}$	$0.344^{\pm 0.039}$	$1.20^{\pm 0.033}$	$0.25^{\pm 0.03}$
Kinetic modeling: local progression and independent tokenizer control								
LCM only	$7.711^{\pm 0.48}$	$0.632^{\pm 0.05}$	$2.769^{\pm 0.15}$	$6.46^{\pm 0.44}$	$0.52^{\pm 0.04}$	$0.542^{\pm 0.06}$	$2.80^{\pm 0.062}$	$0.52^{\pm 0.06}$
LCM + SA	$5.083^{\pm 0.24}$	$0.425^{\pm 0.06}$	$2.728^{\pm 0.13}$	$\mathbf{2.49^{\pm 0.23}}$	$0.71^{\pm 0.035}$	$0.548^{\pm 0.054}$	$2.82^{\pm 0.067}$	$0.66^{\pm 0.06}$
LCM + Token-only	$9.212^{\pm 0.26}$	$0.907^{\pm 0.08}$	$3.009^{\pm 0.19}$	$5.90^{\pm 0.31}$	$0.80^{\pm 0.034}$	$0.647^{\pm 0.065}$	$1.85^{\pm 0.024}$	$0.24^{\pm 0.09}$
LCM + SA + M	$7.766^{\pm 0.34}$	$0.610^{\pm 0.07}$	$2.937^{\pm 0.13}$	$2.54^{\pm 0.26}$	$0.73^{\pm 0.042}$	$0.361^{\pm 0.057}$	$2.35^{\pm 0.043}$	$0.48^{\pm 0.05}$
LCM + SA + M + C	$7.572^{\pm 0.27}$	$0.377^{\pm 0.04}$	$2.936^{\pm 0.13}$	$2.55^{\pm 0.38}$	$0.71^{\pm 0.033}$	$0.451^{\pm 0.052}$	$1.35^{\pm 0.026}$	$0.20^{\pm 0.06}$
LCM + SA + M + C + T	$4.471^{\pm 0.23}$	$0.434^{\pm 0.09}$	$3.000^{\pm 0.15}$	$3.49^{\pm 0.47}$	$0.84^{\pm 0.043}$	$0.119^{\pm 0.011}$	$1.30^{\pm 0.020}$	$0.17^{\pm 0.09}$
Final multi-lag configuration: entropy-calibration comparison								
EntroKine3D w/o entropy calibration	$5.072^{\pm 0.12}$	$0.410^{\pm 0.09}$	$2.914^{\pm 0.08}$	$2.57^{\pm 0.25}$	$0.84^{\pm 0.07}$	$\mathbf{0.118^{\pm 0.03}}$	$1.04^{\pm 0.023}$	$0.18^{\pm 0.03}$
EntroKine3D	$\mathbf{4.342^{\pm 0.22}}$	$\mathbf{0.398^{\pm 0.03}}$	$\mathbf{3.072^{\pm 0.12}}$	$2.65^{\pm 0.21}$	$\mathbf{0.89^{\pm 0.03}}$	$0.122^{\pm 0.02}$	$0.90^{\pm 0.017}$	$\mathbf{0.08^{\pm 0.06}}$

(Table 2). Fixed-condition NS further increases from T0 to entropy-free and Full (Appendix D.9, Table 25), showing that structured profile supervision broadens the stochastic response beyond the observed realization.



Figure 3: **Three direct samples under one condition.** Music, dance, checkpoint, sampler, renderer, and timestamps are fixed. Each seed determines the initial latent and intermediate Gaussian draws. Changes in framing demonstrate stochastic response; Table 17 separately characterizes noise spread and kinetic diagnostics.

Temporal organization complements occupancy supervision. The local progression in Table 2 is non-monotonic: occupancy terms mainly reshape state usage, whereas adding temporal supervision substantially improves sequence quality and continuity. This indicates that broader state coverage is useful only when those states are organized over time. The final model uses the separate multi-lag configuration in Appendix D.1; non-overlapping, block-permutation, and jerk controls further separate temporal ordering from state frequency and local smoothness (Appendix D.8).

Condition correspondence and temporal order. Condition shifting or shuffling degrades conditional agreement, and using only the coarse Y state underperforms the full music–dance condition. Separately, state permutation preserves occupancy while strongly worsening transition statistics, confirming that occurrence and order capture different structure (Appendices D.7 and D.8).

Uncertainty calibration. Full and entropy-free share the same profile discrepancies. Adding calibration improves FID-K and moves DIV-K closer to GT while preserving continuity (Table 2). A gradient-matched entropy-free control narrows but does not close the gap, indicating that the gain is not explained by auxiliary-gradient magnitude alone (Appendix D.4); complete uncertainty diagnostics are in Appendix D.3.

Table 3: Fixed-condition stochastic response (NS) jointly read with task-level quality (FID-K) and geometric validity (Invalid). Protocol and values identical to the complete statistics in Table 25 (Appendix D.9). $a^{\pm b}$: checkpoint SD; $a^{\pm b \times 10^{-c}}$: paired-bootstrap interval.

Method	NS ↑	FID-K ↓	Invalid ↓
DanceCamera3D (Wang et al., 2024a)	N/A (det.)	9.555	0.006
T0: Paired	$0.08^{\pm 0.02}$	$6.381^{\pm 0.13}$	$0.0030^{\pm 3\times 10^{-4}}$
LCM w/o E-KPM	$0.14^{\pm 0.02}$	$7.711^{\pm 0.48}$	$0.0057^{\pm 1\times 10^{-3}}$
w/o entropy calibration	$0.26^{\pm 0.04}$	$5.072^{\pm 0.12}$	$0.0030^{\pm 1\times 10^{-4}}$
EntroKine3D	$\mathbf{0.34^{\pm 0.03}}$	$\mathbf{4.342^{\pm 0.22}}$	$\mathbf{0.0021^{\pm 4\times 10^{-4}}}$

Table 4: Same-condition three-sample set evaluation on a 1–7 scale (18 disjoint test conditions, three unfiltered sampling streams per set). Brackets give cluster-bootstrap 95% intervals; the full protocol is in Appendix F.

Method	Naturalness ↑	Condition consistency ↑	Effective diversity ↑
EntroKine3D	**6.2 [6.05,6.34]**	**6.4 [6.26,6.53]**	**6.5 [6.36,6.63]**
Entropy-free	6.0 [5.85,6.14]	5.8 [5.64,5.95]	5.6 [5.43,5.76]
T0 Paired	5.7 [5.54,5.85]	5.9 [5.75,6.04]	3.0 [2.82,3.18]
LCM w/o E-KPM	4.8 [4.63,4.97]	4.5 [4.32,4.68]	4.6 [4.42,4.78]

Same-condition alternatives. Global diversity does not reveal whether multiple samples from one performance are meaningful alternatives. Figure 3 therefore fixes the condition and varies only the sampling stream. Table 3 shows that Full substantially broadens the fixed-condition response while retaining the best FID-K and lowest invalid-view rate among the stochastic controls. The complete protocol is in Appendix D.9, and the same-condition human study independently confirms higher effective diversity together with strong naturalness and condition consistency (Section 4.4, Table 4).

Efficiency. Four evaluations require $7.26^{\pm 0.14}$ ms versus 68.19 ms for the teacher, a $9.39\times$ speedup; the complete NFE–quality trade-off is reported in Appendix B, Table 6. The corrected rollout keeps $\mathrm{TransKL}_5$ near training-path predictions and below the legacy sampler (Appendix B, Table 7).

4.4 Human Evaluation

A seven-point study with 2,520 presentations, 48 participants, 30 conditions, and seven methods gives Full the highest overall, naturalness, responsiveness, and framing means ($6.6/6.1/6.4/6.7$), with an odds ratio of 2.05 versus entropy-free (Appendix F, Tables 28 and 29); Appendix E reports the expert assessment.

The same-condition three-sample evaluation tests whether stochastic alternatives remain usable views of one performance. Full attains the highest naturalness (6.2), condition consistency (6.4), and effective diversity (6.5); removing entropy calibration reduces effective diversity to 5.6, while paired supervision collapses it to 3.0, indicating that the added variation remains distinct and plausible.

5 Conclusion

Single-reference camera data provide only one accepted realization per performance, leaving plausible alternatives unobserved. EntroKine3D supplements this supervision with shared kinetic-profile constraints over camera-motion geometry, condition-dependent allocation, and temporal evolution, pooled from related performance states and uncertainty-calibrated under finite optimization. Controlled comparisons show that instance-level targets yield narrow fixed-condition responses, state coverage alone is insufficient without temporal organization, and uncertainty calibration improves the quality–variation balance beyond matched gradient budgets. Fixed-condition sampling and same-condition human evaluation show that the resulting alternatives are distinct yet plausible. Combined with a four-step latent generator, shared kinetic-profile supervision enables varied, high-fidelity dance camerawork without sacrificing stochastic flexibility or efficiency.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting, language polishing, and method implementation, and provided feedback on experimental design and interpretation. All AI-assisted work was reviewed by the authors, and the experimental results, analyses, and scientific claims were verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

The appendix provides the data processing, model architecture, training, retrieval, and inference details needed to reproduce DanceMAR. The experimental setup specifies the dataset split, evaluation metrics, and implementation settings used for the main results, while the additional experiments document the corresponding ablation and cross-dataset protocols.

REPRODUCIBILITY STATEMENT

Appendix G specifies configuration selection, uncertainty reporting, and runtime protocols; Appendix A gives the E-KPM construction and training hyperparameters; Appendix F describes the human-evaluation protocol. Training statistics, codebooks, and the human-study manifest will be released with the paper.

REFERENCES

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

Marc Christie, Patrick Olivier, and Jean-Marie Normand. Camera control in computer graphics. *Computer Graphics Forum*, 27(8):2197–2218, 2008. doi: 10.1111/j.1467-8659.2008.01181.x.

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. E.T. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In *European Conference on Computer Vision*, pp. 464–480. Springer, 2024.

Robin Courant, Xi Wang, David Loiseaux, Marc Christie, and Vicky Kalogeiton. Pulp motion: Framing-aware multimodal camera and human motion generation. In *International Conference on Learning Representations*, 2026.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Zahra Dehghanian, Pouya Ardekhani, Amir Vahedi, Hamid Beigy, and Hamid R. Rabiee. Camera trajectory generation: A comprehensive survey of methods, metrics, and future directions. *arXiv preprint arXiv:2506.00974*, 2025.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1574–1583, 2019.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *International Conference on Learning Representations*, 2021.

Xiaoying Huang, Sanyi Zhang, Xirui Wang, Qin Zhang, and Long Ye. DanceStyleCam: Style-based 3d multi-style dance camera movement synthesis. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-ECAI 2026)*, 2026. To appear.

Hongda Jiang, Bin Wang, Xi Wang, Marc Christie, and Baoquan Chen. Example-driven virtual cinematography by learning camera behaviors. *ACM Transactions on Graphics*, 39(4):45:1–45:14, 2020. doi: 10.1145/3386569.3392427.

Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model. *Computer Graphics Forum*, 43(2):e15055, 2024. doi: 10.1111/cgf.15055.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10234–10243, 2023.

Ronghui Li, Yuxiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse-to-fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1524–1534, 2024a.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3d dance generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13381–13392, 2021.

Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation from text. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, 1932.

Christophe Lino and Marc Christie. Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics*, 34(4):82:1–82:12, 2015. doi: 10.1145/2766965.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. In *International Conference on Learning Representations*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic GPT with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11050–11059, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265, 2015.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 32211–32252, 2023.

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.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. EDGE: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–458, 2023.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zixuan Wang, Jia Jia, Shikun Sun, Haozhe Wu, Rong Han, Zhenyu Li, Di Tang, Jiaqing Zhou, and Jiebo Luo. Dancecamera3d: 3d camera movement synthesis with music and dance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7892–7901, 2024a.

Zixuan Wang, Jiayi Li, Xiaoyu Qin, Shikun Sun, Songtao Zhou, Jia Jia, and Jiebo Luo. Dancecamanimator: Keyframe-based controllable 3d dance camera synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024b. doi: 10.1145/3664647.3680980.

Chenghao Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. Smooth and flexible camera movement synthesis via temporal masked generative modeling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 14816–14834, 2025. doi: 10.52202/085713-0446.

Jinming Zhang, Yunlian Sun, Hongwen Zhang, and Jinhui Tang. EDMG: Towards efficient long dance motion generation with fundamental movements from dance genres. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10447–10456, 2025a. doi: 10.1145/3746027.3755677.

Mengchen Zhang, Tong Wu, Jing Tan, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. Gendop: Auto-regressive camera trajectory generation as a director of photography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b. doi: 10.1109/ICCV51701.2025.01694.

Appendix

A E-KPM Construction and Estimation Details

This appendix gives the complete construction of the kinetic descriptors, performance states, discrete kinetic profiles, temporal joints, and uncertainty statistics used by E-KPM. All fitted transforms and reference statistics are obtained from the training split only and remain fixed during optimization and testing.

A.1 Kinetic Descriptors and Performance States

Window-level camera kinetics. The camera descriptor has five channels $(\mathbf{e}, \mathbf{b}^z, \mathbf{b}^y, \mathbf{b}^x, f)$ and two temporal statistics for each channel, giving the ten-dimensional descriptor in Eq. equation 2. For a camera channel $\mathbf{x}$ and a window containing n_i frames, let

$$\mathcal{V}_j = \{h : 1 \le h < n_i,\ |h-j| \le r_s\}, \qquad \mathcal{A}_j = \{h : 1 \le h < n_i - 1,\ |h-j| \le r_s\},$$

with smoothing radius $r_s = 2$. We define locally averaged first- and second-order differences

$$\overline{\mathbf{v}}_j(\mathbf{x}) = \frac{1}{|\mathcal{V}_j|} \sum_{h \in \mathcal{V}_j} \frac{\mathbf{x}_{h+1} - \mathbf{x}_h}{\Delta t}, \tag{18}$$

$$\overline{\mathbf{a}}_j(\mathbf{x}) = \frac{1}{|\mathcal{A}_j|} \sum_{h \in \mathcal{A}_j} \frac{\mathbf{x}_{h+2} - 2\mathbf{x}_{h+1} + \mathbf{x}_h}{\Delta t^2}. \tag{19}$$

The window statistics are

$$V_i(\mathbf{x}) = \frac{1}{n_i - 1} \sum_{j=1}^{n_i-1} \|\overline{\mathbf{v}}_j(\mathbf{x})\|_2^2, \qquad A_i(\mathbf{x}) = \frac{1}{n_i - 2} \sum_{j=1}^{n_i-2} \|\overline{\mathbf{a}}_j(\mathbf{x})\|_2. \tag{20}$$

We use $\Delta t = 1/30$ seconds. Only valid differences are included. Camera-channel normalization is fitted on training windows and shared by reference and generated descriptors.

Music–dance performance descriptor. Let

$$u_t = \log(\epsilon_a + \|\mathbf{a}_t\|_2), \qquad o_t = \|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2 \tag{21}$$

denote the audio log-energy and onset proxy, respectively, with $o_1 = 0$. For dancer joint j, let $\mathbf{p}_{t,j}$ denote its position and $\mathbf{r}_{t,j} = \mathbf{p}_{t,j} - \mathbf{p}_{t,1}$ its root-relative coordinate. Let

$$h_t = \max_j r_{t,j}^{(y)} - \min_j r_{t,j}^{(y)}$$

be the dancer height scale. We define the normalized motion activity

$$s_t = \frac{1}{J \max(h_t, \epsilon_h)} \sum_{j=1}^{J} \|\mathbf{r}_{t,j} - \mathbf{r}_{t-1,j}\|_2, \qquad t \ge 2, \tag{22}$$

with $s_1 = 0$. For window i,

$$\Phi_Y(A, M)_i = [\mu_i(u), \sigma_i(u), \mu_i(o), \sigma_i(o), \mu_i(s), \sigma_i(s)]. \tag{23}$$

The six-dimensional descriptors are standardized using training statistics. K-means++ is fitted on training windows to obtain $\{\boldsymbol{\nu}_y\}_{y=1}^{N_Y}$, giving

$$Y_i = \arg \min_{y \in \{1,\dots,N_Y\}} \left\| \widetilde{\Phi}_Y(A, M)_i - \boldsymbol{\nu}_y \right\|_2^2. \tag{24}$$

We use $N_Y = 16$. The full generator condition $E_{\mathrm{cond}}(A, M)$ is separate from this coarse profiling state.

A.2 Continuous Alignment and Soft Kinetic States

The continuous kinetic discrepancy is defined in Eq. equation 7. We fit $N_\kappa = 32$ K-means++ kinetic centers $\{\boldsymbol{\mu}_k\}_{k=1}^{N_\kappa}$ on standardized training descriptors and keep this codebook fixed. Reference and generated windows use the same centers, and their soft state assignments follow Eq. equation 4 with $\tau = 1$.

A.3 All-Window Occupancy Profiles

Let $\mathcal{W}_d(y)$ denote the valid windows assigned to performance state y, where $d \in \{r, g\}$ denotes reference or generated data, and $n_d(y) = |\mathcal{W}_d(y)|$. With probability floor δ, the conditional and marginal occupancy profiles are

$$p_d(k \mid y) = \frac{\sum_{i \in \mathcal{W}_d(y)} q_i^d(k) + \delta}{n_d(y) + N_\kappa \delta}, \tag{25}$$

$$p_d(k) = \frac{\sum_i q_i^d(k) + \delta}{|\mathcal{W}_d| + N_\kappa \delta}. \tag{26}$$

The conditional occupancy weights are fixed from the training data:

$$\omega_y^{\mathrm{occ}} = \frac{n_r(y)}{\sum_v n_r(v)}. \tag{27}$$

These occupancy estimates are used in the marginal and condition-dependent discrepancies defined in Eq. equation 9.

A.4 Past-Conditioned Temporal Profiles

For the final temporal profile, the source kinetic window contains $W = 75$ frames. For source start s_i, the past condition state is computed from the preceding music–dance window:

$$\mathbf{K}_{d,i}^- = \mathcal{K}\left(C_{s_i:s_i+W-1}^d\right), \qquad Y_i^- = Q_Y\left(\Phi_Y(A_{s_i-W:s_i-1}, M_{s_i-W:s_i-1})\right). \tag{28}$$

The future camera descriptor at lag ℓ is

$$\mathbf{K}_{d,i,\ell}^+ = \mathcal{K}\left(C_{s_i+\ell:s_i+\ell+W-1}^d\right), \qquad \ell \in \Lambda_{\mathrm{TE}} = \{75, 150\}. \tag{29}$$

Source starts are spaced 75 frames apart. The past condition, source camera, and future camera supports are non-overlapping.

For sequence length T, the eligible temporal anchors are

$$\mathcal{A}_{d,\ell}(T) = \{s_i = s_0 + 75j : j \geq 0,\ s_i - W \geq 0,\ s_i + W - 1 < T,\ s_i + \ell + W - 1 < T\}. \tag{30}$$

Writing $q_{d,i}^- = q(\mathbf{K}_{d,i}^-)$ and $q_{d,i,\ell}^+ = q(\mathbf{K}_{d,i,\ell}^+)$, the soft temporal counts are

$$N_{d,\ell}(y, k, k') = \sum_{i \in \mathcal{A}_{d,\ell}} \mathbb{1}[Y_i^- = y]\, q_{d,i}^-(k)\, q_{d,i,\ell}^+(k'). \tag{31}$$

Let

$$n_{y,\ell}^r = \sum_{k,k'} N_{r,\ell}(y, k, k')$$

and define the represented support

$$\mathcal{S}_\ell = \{y : n_{y,\ell}^r \geq n_{\min}\}, \qquad n_{\min} = 4.$$

The frozen train-set pair mixture is

$$\omega_{y,\ell}^{\mathrm{pair}} = \frac{n_{y,\ell}^r}{\sum_{v \in \mathcal{S}_\ell} n_{v,\ell}^r}, \qquad y \in \mathcal{S}_\ell. \tag{32}$$

The reference joint and generated minibatch joint are

$$J^{\star}_{r,\ell}(y,k,k') = \omega^{\mathrm{pair}}_{y,\ell} \frac{N^{\star}_{r,\ell}(y,k,k') + \delta}{n^{r}_{y,\ell} + N^{2}_{\kappa}\delta}, \tag{33}$$

$$\widehat{J}^{B}_{g,\ell}(y,k,k') = \omega^{\mathrm{pair}}_{y,\ell} \frac{\widehat{N}^{B}_{g,\ell}(y,k,k') + \delta}{b_{\mathrm{pair}} + N^{2}_{\kappa}\delta}. \tag{34}$$

We use $b_{\mathrm{pair}} = 2$ generated anchors for each active (y,ℓ) stratum. The reference joint, support, and pair weights are frozen once from the training split.

Marginalization of $J_{d,\ell}$ gives the source-state occupancy $\pi_{d,\ell}(k)$, pair occupancy $\pi_{d,\ell}(y,k)$, the transition $T_{d,\ell}(k' \mid k)$, and the performance-conditioned transition $T^{Y^-}_{d,\ell}(k' \mid k, y)$. These distributions are used in the temporal discrepancies $\mathcal{D}_T$ and $\mathcal{D}_{CT}$ defined in Eq. equation 12.

A.5 Uncertainty Statistics and Factorization

Using natural logarithms and the convention $0 \log 0 = 0$, the marginal and conditional occupancy-entropy gaps are

$$\Delta H_M = \left| H[p_r(k)] - H[p_g(k)] \right|, \tag{35}$$

$$\Delta H_C = \sum_y \omega^{\mathrm{occ}}_y \left| H[p_r(k \mid y)] - H[p_g(k \mid y)] \right|. \tag{36}$$

For the temporal joint,

$$H^{d}_{\mathrm{tr},\ell} = \sum_k \pi_{d,\ell}(k) H[T_{d,\ell}(\cdot \mid k)], \tag{37}$$

$$H^{d}_{\mathrm{cond},\ell} = \sum_{y,k} \pi_{d,\ell}(y,k) H[T^{Y^-}_{d,\ell}(\cdot \mid k, y)]. \tag{38}$$

The lag-weighted entropy gaps are

$$\Delta H_{\mathrm{tr}} = \sum_{\ell \in \Lambda_{\mathrm{TE}}} \omega_\ell \left| H^{r}_{\mathrm{tr},\ell} - H^{g}_{\mathrm{tr},\ell} \right|, \tag{39}$$

$$\Delta H_{\mathrm{cond}} = \sum_{\ell \in \Lambda_{\mathrm{TE}}} \omega_\ell \left| H^{r}_{\mathrm{cond},\ell} - H^{g}_{\mathrm{cond},\ell} \right|. \tag{40}$$

The finite-lag conditional-information statistic is

$$\mathrm{TE}_{d,\ell} = \sum_{y,k,k'} J_{d,\ell}(y,k,k') \log \frac{T^{Y^-}_{d,\ell}(k' \mid k, y)}{T_{d,\ell}(k' \mid k)}. \tag{41}$$

Equivalently,

$$\mathrm{TE}_{d,\ell} = H^{d}_{\mathrm{tr},\ell} - H^{d}_{\mathrm{cond},\ell}. \tag{42}$$

Its reference gap is

$$\Delta_{\mathrm{TE}} = \sum_{\ell \in \Lambda_{\mathrm{TE}}} \omega_\ell \left| \mathrm{TE}_{r,\ell} - \mathrm{TE}_{g,\ell} \right|. \tag{43}$$

Each lag-specific joint factors as

$$J_{d,\ell}(y,k,k') = \omega^{\mathrm{pair}}_{y,\ell} \pi_{d,\ell}(k \mid y) T^{Y^-}_{d,\ell}(k' \mid k, y). \tag{44}$$

Consequently,

$$D_{\mathrm{KL}}(J_{r,\ell} \| J_{g,\ell}) = \sum_y \omega^{\mathrm{pair}}_{y,\ell} D_{\mathrm{KL}}\left(\pi_{r,\ell}(\cdot \mid y) \| \pi_{g,\ell}(\cdot \mid y)\right) + \mathcal{D}_{CT,\ell}. \tag{45}$$

Let

$$e_{\mathrm{tr},\ell} = H^{r}_{\mathrm{tr},\ell} - H^{g}_{\mathrm{tr},\ell}, \qquad e_{\mathrm{cond},\ell} = H^{r}_{\mathrm{cond},\ell} - H^{g}_{\mathrm{cond},\ell}.$$

Table 5: Training and kinetic-profile hyperparameters selected on the validation split.

Parameter	Value	Parameter	Value
N_Y	16	N_κ	32
τ	1.0	window-start stride	5 frames
W_{Y^-}, W_K, W_{K^+}	75,75,75	Λ_{TE}	$\{75, 150\}$
$\omega_{75}, \omega_{150}$	0.5,0.5	$b_{\mathrm{pair}}, n_{\min}$	2,4
$\varepsilon_{\mathrm{OT}}$	0.10	Sinkhorn iterations	50
λ_{SA}	1.0	λ_M, λ_C	0.25,0.25
λ_T, λ_{CT}	0.15,0.15	γ_M, γ_C	0.05,0.05
$\gamma_{\mathrm{tr}}, \gamma_{CT}, \gamma_{\mathrm{TE}}$	0.05,0.05,0.10	δ	10^{-8}

Then

$$\Delta_{\mathrm{TE}} = \sum_{\ell} \omega_\ell \left| e_{\mathrm{tr},\ell} - e_{\mathrm{cond},\ell} \right| \leq \Delta H_{\mathrm{tr}} + \Delta H_{\mathrm{cond}}. \tag{46}$$

These statistics are differentiable plug-in estimates computed from the same soft temporal joint. The finite-lag quantity measures conditional dependence under the defined history and lag, not interventional causality.

A.6 Training Procedure

For every training update, we first sample an ordered camera block and its aligned music–dance condition and construct the noisy camera latent. The online student produces a clean latent and the frozen teacher and EMA student provide the latent-space training targets.

The online clean prediction is decoded with the frozen camera VAE. All valid camera windows are then converted into kinetic descriptors for continuous alignment and occupancy supervision. For every $\ell \in \Lambda_{\mathrm{TE}}$, generated temporal anchors are sampled within their active train-set strata and accumulated into $\widehat{J}^B_{g,\ell}$. The profile and uncertainty losses are computed against the frozen training references.

Only the online student parameters θ are optimized. The teacher, VAE, performance-state centroids, kinetic codebook, reference supports, and reference profiles are fixed. The EMA student is updated after each optimizer step.

A.7 E-KPM Hyperparameters

B Few-Step Latent Camera Generation and Sampling

This appendix specifies the latent diffusion teacher, student distillation objective, EMA consistency target, and four-step variance-corrected sampling process used by Section 3.2 and Section 3.4.

B.1 Forward Noising and Latent Distillation

The camera VAE encodes an authored trajectory into

$$\mathbf{z}_0 = E_{\mathrm{vae}}(C).$$

We use a 1,000-index linear diffusion schedule $t \in \{0, \ldots, 999\}$ with

$$\beta_t = 10^{-4} + \frac{t}{999}(0.02 - 10^{-4}), \qquad \bar{\alpha}_t = \prod_{u=0}^{t} (1 - \beta_u). \tag{47}$$

Forward noising gives

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{48}$$

A frozen conditional diffusion teacher F_T provides the clean-latent target:

$$\mathcal{L}_{x_0} = \left\| F_\theta(\mathbf{z}_t, t, \mathbf{c}) - \mathrm{sg}[F_T(\mathbf{z}_t, t, \mathbf{c})] \right\|_2^2. \tag{49}$$

For

$$\widehat{\mathbf{z}}_{0,t}^{T} = F_T(\mathbf{z}_t, t, \mathbf{c}),$$

the teacher-implied noise is

$$\widehat{\boldsymbol{\epsilon}}_t^{T} = \frac{\mathbf{z}_t - \sqrt{\bar{\alpha}_t}\widehat{\mathbf{z}}_{0,t}^{T}}{\sqrt{1-\bar{\alpha}_t}}. \tag{50}$$

For a lower noise level $s < t$, the teacher transport is

$$\mathbf{z}_s^{T} = \sqrt{\bar{\alpha}_s}\widehat{\mathbf{z}}_{0,t}^{T} + \sqrt{1-\bar{\alpha}_s}\widehat{\boldsymbol{\epsilon}}_t^{T}. \tag{51}$$

An exponential-moving-average student $F_{\bar{\theta}}$ provides the adjacent-level consistency target:

$$\mathcal{L}_{\text{cons}} = \left\| F_\theta(\mathbf{z}_t, t, \mathbf{c}) - \text{sg}[F_{\bar{\theta}}(\mathbf{z}_s^{T}, s, \mathbf{c})] \right\|_2^2. \tag{52}$$

The target parameters are updated as

$$\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta, \qquad \mu = 0.999. \tag{53}$$

Only the online student receives gradients.

The base latent objective is

$$\mathcal{L}_{\text{base}} = \lambda_{x_0}\mathcal{L}_{x_0} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}, \qquad \lambda_{x_0} = 1, \quad \lambda_{\text{cons}} = 10^{-3}. \tag{54}$$

B.2 Variance-Corrected Stochastic Sampling

At inference, the teacher prediction is replaced with the online student prediction

$$\widehat{\mathbf{z}}_{0,t}^{S} = F_\theta(\mathbf{z}_t, t, \mathbf{c}). \tag{55}$$

Its implied noise is

$$\widehat{\boldsymbol{\epsilon}}_t^{S} = \frac{\mathbf{z}_t - \sqrt{\bar{\alpha}_t}\widehat{\mathbf{z}}_{0,t}^{S}}{\sqrt{1-\bar{\alpha}_t}}. \tag{56}$$

For adjacent inference levels $t > s$, we use

$$\sigma_{t,s}(\eta) = \eta\sqrt{\frac{1-\bar{\alpha}_s}{1-\bar{\alpha}_t}}\sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_s}}, \tag{57}$$

and propagate the latent as

$$\begin{aligned}\mathbf{z}_s = \sqrt{\bar{\alpha}_s}\widehat{\mathbf{z}}_{0,t}^{S} + \sqrt{1-\bar{\alpha}_s-\sigma_{t,s}^2}\widehat{\boldsymbol{\epsilon}}_t^{S} \\ + \sigma_{t,s}\boldsymbol{\xi}_s, \qquad \boldsymbol{\xi}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).\end{aligned} \tag{58}$$

The squared coefficients of the two stochastic components sum to $1-\bar{\alpha}_s$, preserving the intended noise variance under the standardized residual interpretation.

For N network evaluations, the inference grid is

$$(t_0, \ldots, t_{N-1}) = \text{long}\left(\text{linspace}(999, 0, N)\right). \tag{59}$$

The selected four-step model therefore evaluates the student at

$$(999, 666, 333, 0). \tag{60}$$

The final network output at $t = 0$ is returned directly as the clean camera latent; no additional transport is performed afterward.

We use $\eta = 0.5$ for the reported model. A complete sampling stream consists of the initial Gaussian latent and all intermediate Gaussian draws $\{\boldsymbol{\xi}_s\}$. Holding the music–dance condition fixed while changing this stream produces the multiple direct samples evaluated in the main paper.

Table 6: Quality–sampling-cost trade-off under the grouped test protocol using the selected variance-corrected sampler. Superscripts are standard deviations over three checkpoint means ($a^{\pm b}$).

Setting	FID-K ↓	SK ↓	DIV-K ≈GT	Lat. (ms) ↓	Speedup ↑
Teacher-50	7.546	8.01	2.697	68.19	1.00×
Student-1	$8.906^{\pm 0.48}$	$6.91^{\pm 0.42}$	$2.383^{\pm 0.12}$	$\mathbf{2.24^{\pm 0.05}}$	**30.44×**
Student-2	$5.327^{\pm 0.31}$	$4.57^{\pm 0.31}$	$2.712^{\pm 0.11}$	$3.86^{\pm 0.08}$	17.67×
Student-4	$4.342^{\pm 0.22}$	$2.65^{\pm 0.21}$	$3.072^{\pm 0.12}$	$7.26^{\pm 0.14}$	9.39×
Student-8	$\mathbf{4.151^{\pm 0.21}}$	$\mathbf{2.50^{\pm 0.18}}$	$\mathbf{3.102^{\pm 0.12}}$	$13.07^{\pm 0.25}$	5.22×

Table 7: Kinetic-profile diagnostics for decoded training predictions and four-step rollouts. Values are mean ± raw-sequence paired-bootstrap 95% intervals.

Path	SK↓	Coverage↑	C.KL↓	TransKL$_5$ ↓	$\Delta H_{\mathrm{tr},5}$	TE.g↓
Training-path prediction	$2.38^{\pm 0.18}$	$0.92^{\pm 0.04}$	$0.81^{\pm 0.05}$	$0.11^{\pm 0.03}$	$0.04^{\pm 0.02}$	$0.05^{\pm 0.02}$
Four-step rollout	$2.65^{\pm 0.21}$	$0.89^{\pm 0.05}$	$0.90^{\pm 0.05}$	$0.15^{\pm 0.03}$	$0.07^{\pm 0.02}$	$0.08^{\pm 0.02}$
Legacy additive rollout	$3.18^{\pm 0.27}$	$0.92^{\pm 0.05}$	$1.23^{\pm 0.10}$	$0.48^{\pm 0.06}$	$0.24^{\pm 0.04}$	$0.21^{\pm 0.04}$

B.3 Sampling Cost and Quality Trade-off

B.4 Training and Deployment Paths

E-KPM is applied during training to decoded clean predictions

$$\widehat{C}_t = D_{\mathrm{vae}}\left(F_\theta(\mathbf{z}_t, t, \mathbf{c})\right)$$

obtained from noised training latents. Deployment instead starts from a Gaussian latent and follows the complete four-step stochastic rollout in Eq. equation 58.

The same frozen VAE decoder is used in both paths. After decoding, deployment additionally applies the deterministic geometry-safe camera projection described in Appendix C.

The four-step rollout remains close to the decoded training prediction on the profile diagnostics, whereas the legacy additive sampler produces a larger temporal-profile shift.

C Evaluation Metrics and Protocols

Evaluation roles. FID-K/FID-S and DIV-K/DIV-S retain the DCM feature-based evaluation protocol; DMR/LCD remain the reported dancer-fidelity measures (Wang et al., 2024a;b). FID compares dataset feature distributions, not per-sample prediction accuracy (Heusel et al., 2017). SK applies the continuous discrepancy in Eq. equation 7 at evaluation. TokenKL and C.KL use $\mathcal{D}_M$ and $\mathcal{D}_C$. $\mathrm{TransKL}_\ell$ is the transition KL at the shown raw-frame lag; $\mathrm{TransKL}_\Lambda = \sum_{\ell\in\Lambda_{\mathrm{TE}}} \omega_\ell \mathrm{TransKL}_\ell$ is the final lag-weighted form of $\mathcal{D}_T$. Unsuffixed TE.g is Δ_{TE} over $\{75, 150\}$; TE.g$_\ell$ denotes a single-lag gap. Local and multi-lag reporting are distinguished below. Training references are fitted only on training data. Held-out statistics evaluate outputs rather than update the generator or its codebooks.

Active-state overlap. Set $\rho = 0.5/N_\kappa$ and $\mathcal{A}_r = \{k : p_r(k) \geq \rho\}$. Coverage is

$$\mathrm{Coverage} = \frac{1}{|\mathcal{A}_r|} \sum_{k\in\mathcal{A}_r} \mathbb{1}[p_g(k) \geq \rho]. \tag{61}$$

Coverage records how many reference-active discrete states are also active in generated samples, under the shared codebook and threshold.

Temporal variation. Let s_{pos} and s_{rot} be the validation-set median scales used by the source evaluator. Define

$$J_{\mathrm{pos}} = \frac{1}{T-2}\sum_{\ell=2}^{T-1}\frac{\|\mathbf{e}_{\ell+1} - 2\mathbf{e}_{\ell} + \mathbf{e}_{\ell-1}\|_2}{s_{\mathrm{pos}}}, \tag{62}$$

$$J_{\mathrm{rot}} = \frac{1}{T-2}\sum_{\ell=2}^{T-1}\frac{d_{SO(3)}(\mathbf{R}_{\ell+1}\mathbf{R}_{\ell}^{-1}, \mathbf{R}_{\ell}\mathbf{R}_{\ell-1}^{-1})}{s_{\mathrm{rot}}}, \qquad \mathrm{CJ} = \tfrac{1}{2}(J_{\mathrm{pos}} + J_{\mathrm{rot}}). \tag{63}$$

$d_{SO(3)}$ denotes angular distance between rotations. CJ summarizes local second-order variation; visibility and human ratings provide complementary assessments of the resulting shots, including intentional cuts.

Noise and condition response. For R samples under original music–dance condition u, let $\mathbf{K}_{u,r}$ denote the kinetic representation used by the source distance d_K. The fixed-condition response is

$$\mathrm{NS}(u) = \frac{2}{R(R-1)}\sum_{r<s}\|\mathbf{K}_{u,r} - \mathbf{K}_{u,s}\|_2. \tag{64}$$

CC.D uses cross-condition pairs under the source evaluation rule; it differs from fixed-condition NS and training-state occupancy. Condition sensitivity compares matched-duration conditions with a shared seed against different seeds under one condition:

$$\mathrm{CS} = \frac{\mathbb{E}[d_K(G(A_a, M_a, \epsilon), G(A_b, M_b, \epsilon))]}{\mathbb{E}[d_K(G(A, M, \epsilon_a), G(A, M, \epsilon_b))] + 10^{-8}}. \tag{65}$$

Here ϵ denotes the complete sampling stream, including intermediate noise. CS is interpreted jointly with its within-condition denominator NS and the conditional discrepancies. The same trajectory representation and distance are used across NS comparisons.

Temporal units and metric names. Window length, start stride, and prediction lag are different quantities. Occupancy windows start every five raw frames; final temporal anchors start every 75 frames; future-camera windows begin 75 or 150 frames after the source start. We use $\mathrm{TransKL}_5$ and $\Delta H_{\mathrm{tr},5}$ for the local columns in the order and rollout analyses. Their labeling follows the lag-resolved values in Table 20; these are distinct from the final lag-weighted summaries. TE.g without a lag suffix denotes $\sum_{\ell\in\{75,150\}}\omega_\ell|\,\mathrm{TE}_{r,\ell} - \mathrm{TE}_{g,\ell}\,|$, whereas TE.g$_\ell$ refers to the indicated lag. Absolute differences are formed per lag before averaging. In particular, a local entropy gap such as 0.07 and the final multi-lag gap 0.09 refer to different temporal supports. The T_5 row marked † retains its local-control information summary 0.17; the same local configuration has past-conditioned multi-lag TE.g 0.31 in Table 23. We use the latter table for shared-support temporal comparisons and keep the former as a local diagnostic, without converting its uncertainty estimate.

Reference fitting and evaluation distributions. Training data determine normalizers, condition and kinetic codebooks, shot-state quantizers, and the reference targets used by the losses. Evaluation-only distribution distances compare generated samples with authored cameras from the corresponding evaluation split, using those frozen transforms. This convention applies to FCD and AVD as well as the other held-out distribution diagnostics; evaluation reference statistics do not update the generator or the training targets.

D Additional Ablations and Diagnostics

D.1 Ablation Configurations

The component progression in Table 2 starts from the latent-consistency student and adds the kinetic-profile terms in stages. We write

$$\mathcal{S} = \lambda_{\mathrm{SA}}\mathcal{D}_{\mathrm{SA}}, \tag{66}$$

$$\mathcal{B}_M = \lambda_M\mathcal{D}_M + \gamma_M\Delta H_M, \tag{67}$$

$$\mathcal{B}_C = \lambda_C\mathcal{D}_C + \gamma_C\Delta H_C, \tag{68}$$

$$\mathcal{B}_T = \lambda_T\mathcal{D}_T + \gamma_{\mathrm{tr}}\Delta H_{\mathrm{tr}}, \tag{69}$$

$$\mathcal{B}_{CT} = \lambda_{CT}\mathcal{D}_{CT} + \gamma_{CT}\Delta H_{\mathrm{cond}} + \gamma_{\mathrm{TE}}\Delta_{\mathrm{TE}}. \tag{70}$$

Table 8: Complete metrics for the supervision and E-KPM ablations on the grouped DCM test split. Repeated rows use the same three-checkpoint, three-seed aggregation and safe projection as Table 1. M/C/T_5 form the local composite-block progression after SA; the final Full/entropy-free pair shares non-overlapping multi-lag profiles. Objectives are defined in Appendix D.1. Token-only is independent. † retains the local-control information summary; shared final-lag comparisons are in Table 23.

Variant	FID_k	FID_s	DIV_k	DIV_s	DMR	LCD	SK	Cov.	CJ	C.KL	CS	TE.g
Diffusion Teacher	7.546	0.603	2.697	1.32	0.0053	0.264	8.01	0.53	0.581	3.00	0.45	0.70
LCM only	$7.711^{\pm 0.48}$	$0.632^{\pm 0.05}$	$2.769^{\pm 0.15}$	$1.54^{\pm 0.11}$	$0.0045^{\pm 6\times 10^{-4}}$	$0.219^{\pm 0.01}$	$6.46^{\pm 0.44}$	$0.52^{\pm 0.04}$	$0.542^{\pm 0.06}$	$2.80^{\pm 0.062}$	$0.52^{\pm 0.05}$	$0.52^{\pm 0.06}$
T0: Paired	$6.381^{\pm 0.13}$	$0.429^{\pm 0.06}$	$2.032^{\pm 0.19}$	$1.200^{\pm 0.12}$	$\mathbf{0.0026^{\pm 3.63\times 10^{-4}}}$	$\mathbf{0.175^{\pm 0.02}}$	$4.73^{\pm 0.26}$	$0.49^{\pm 0.025}$	$0.267^{\pm 0.021}$	$2.65^{\pm 0.063}$	$0.30^{\pm 0.01}$	$0.73^{\pm 0.06}$
T1: Global set	$6.408^{\pm 0.38}$	$0.735^{\pm 0.09}$	$2.729^{\pm 0.16}$	$1.503^{\pm 0.14}$	$0.0049^{\pm 9.89\times 10^{-4}}$	$0.238^{\pm 0.01}$	$5.24^{\pm 0.45}$	$0.61^{\pm 0.033}$	$0.712^{\pm 0.058}$	$2.90^{\pm 0.063}$	$0.44^{\pm 0.03}$	$0.66^{\pm 0.07}$
T2: Condition-bin set	$5.486^{\pm 0.25}$	$0.595^{\pm 0.08}$	$2.887^{\pm 0.14}$	$1.562^{\pm 0.17}$	$0.0040^{\pm 3.96\times 10^{-4}}$	$0.221^{\pm 0.02}$	$3.75^{\pm 0.27}$	$0.73^{\pm 0.036}$	$0.481^{\pm 0.042}$	$\mathbf{0.42^{\pm 0.044}}$	$0.86^{\pm 0.09}$	$0.42^{\pm 0.05}$
T3: Set distribution	$4.851^{\pm 0.29}$	$0.477^{\pm 0.05}$	$2.995^{\pm 0.12}$	$1.591^{\pm 0.18}$	$0.0036^{\pm 6.26\times 10^{-4}}$	$0.213^{\pm 0.03}$	$3.15^{\pm 0.21}$	$0.81^{\pm 0.031}$	$0.344^{\pm 0.039}$	$1.20^{\pm 0.033}$	$1.12^{\pm 0.07}$	$0.25^{\pm 0.03}$
LCM + SA	$5.083^{\pm 0.24}$	$0.425^{\pm 0.06}$	$2.728^{\pm 0.13}$	$1.524^{\pm 0.19}$	$0.0045^{\pm 2.72\times 10^{-4}}$	$0.221^{\pm 0.08}$	$\mathbf{2.49^{\pm 0.23}}$	$0.71^{\pm 0.035}$	$0.548^{\pm 0.054}$	$2.82^{\pm 0.067}$	$0.50^{\pm 0.08}$	$0.66^{\pm 0.06}$
LCM + Token-only	$9.212^{\pm 0.26}$	$0.907^{\pm 0.08}$	$3.009^{\pm 0.19}$	$1.623^{\pm 0.13}$	$0.0041^{\pm 6.36\times 10^{-4}}$	$0.213^{\pm 0.07}$	$5.90^{\pm 0.31}$	$0.80^{\pm 0.034}$	$0.647^{\pm 0.065}$	$1.85^{\pm 0.024}$	$0.82^{\pm 0.05}$	$0.24^{\pm 0.09}$
LCM + SA + M	$7.766^{\pm 0.34}$	$0.610^{\pm 0.07}$	$2.937^{\pm 0.13}$	$1.581^{\pm 0.12}$	$0.0049^{\pm 8.58\times 10^{-4}}$	$0.237^{\pm 0.05}$	$2.54^{\pm 0.26}$	$0.73^{\pm 0.042}$	$0.361^{\pm 0.057}$	$2.35^{\pm 0.043}$	$0.65^{\pm 0.05}$	$0.48^{\pm 0.05}$
LCM + SA + M + C	$7.572^{\pm 0.27}$	$0.377^{\pm 0.04}$	$2.936^{\pm 0.13}$	$\mathbf{1.672^{\pm 0.19}}$	$0.0033^{\pm 5.35\times 10^{-4}}$	$0.239^{\pm 0.02}$	$2.55^{\pm 0.38}$	$0.71^{\pm 0.033}$	$0.451^{\pm 0.052}$	$1.35^{\pm 0.026}$	$0.92^{\pm 0.07}$	$0.20^{\pm 0.06}$
LCM + SA + M + C + T	$4.471^{\pm 0.23}$	$0.434^{\pm 0.09}$	$3.000^{\pm 0.15}$	$1.610^{\pm 0.14}$	$0.0036^{\pm 8.78\times 10^{-4}}$	$0.209^{\pm 0.04}$	$3.49^{\pm 0.47}$	$0.84^{\pm 0.043}$	$0.119^{\pm 0.011}$	$1.30^{\pm 0.020}$	$0.88^{\pm 0.05}$	$0.17^{\pm 0.09}$ †
EntroKine3D w/o entropy calibration	$5.072^{\pm 0.12}$	$0.410^{\pm 0.09}$	$2.914^{\pm 0.08}$	$1.59^{\pm 0.14}$	$0.0037^{\pm 6\times 10^{-4}}$	$0.211^{\pm 0.014}$	$2.57^{\pm 0.25}$	$0.84^{\pm 0.07}$	$\mathbf{0.118^{\pm 0.03}}$	$1.04^{\pm 0.023}$	–	$0.18^{\pm 0.03}$
EntroKine3D	$\mathbf{4.342^{\pm 0.22}}$	$\mathbf{0.398^{\pm 0.03}}$	$\mathbf{3.072^{\pm 0.12}}$	$1.62^{\pm 0.08}$	$0.0032^{\pm 4\times 10^{-4}}$	$0.208^{\pm 0.01}$	$2.65^{\pm 0.21}$	$\mathbf{0.89^{\pm 0.03}}$	$0.122^{\pm 0.02}$	$0.90^{\pm 0.017}$	$\mathbf{1.55^{\pm 0.07}}$	$\mathbf{0.08^{\pm 0.06}}$

The local progression uses $\mathcal{S}$, then adds $\mathcal{B}_M$ and $\mathcal{B}_C$. Its last temporal row uses

$$\mathcal{B}_T^{(5)} = \lambda_T \mathcal{D}_{T,5} + \gamma_{\mathrm{tr}} \Delta H_{\mathrm{tr},5}. \tag{71}$$

The final EntroKine3D configuration instead uses the non-overlapping lags $\{75, 150\}$ together with $\mathcal{B}_{CT}$. The entropy-free variant keeps the same profile discrepancies and removes the explicit entropy and information-gap terms.

Table 8 extends the main-text ablation with the full metric set. It separates supervision formulations, the local T_5 progression, and the final multi-lag Full/entropy-free pair.

Quality and statistical structure. T0 retains the lowest mean DMR and LCD, but it is worse than Full on FID-K and FID-S and remains substantially lower in DIV and active-state Coverage. This supports the distinction between point-wise dancer fidelity and distributional quality–variation behavior. T3 improves FID-K, Coverage, CS, and TE.g relative to T2, while T2 retains lower C.KL. These results motivate evaluating conditional allocation together with occupancy and temporal organization.

Cumulative and removal comparisons. SA has the lowest reported SK. The transition and entropy-free configurations have similarly low mean CJ. Full combines lower FID-K with higher Coverage and CS and lower TE.g. The entropy-free row retains all divergences and therefore specifically tests the entropy-calibration branch at the fixed configuration. Direct uncertainty diagnostics follow in Table 10; they describe the uncertainty geometry calibrated by that branch.

D.2 Definitions of the Supervision-Formulation Controls

For the supervision-formulation controls, let $\widehat{C}_i$ denote the student camera associated with condition i. Table 9 defines each control by its reference association and its supervision operator; every variant adds its operator to the shared base objective $\mathcal{L}_{\mathrm{base}}$ of the base objective above and keeps the backbone, representation, renderer, and training route unchanged.

Here ℓ_{cam} denotes the sample-wise camera-space regression cost applied to the decoded camera representation before geometry-safe projection. The set discrepancy $\mathcal{D}_{\mathrm{set}}$ uses the debiased Sinkhorn discrepancy in Eq. equation 7 on the state-conditioned kinetic-feature collection, and ω_y^{set} is the represented reference-state mixture for that comparison. The token objective $\mathcal{R}_{\mathrm{token}}$ is the token-level cross-entropy against the frozen codebook assignments. Token-only therefore tests whether exposing discrete kinetic states alone is sufficient without explicitly matching their empirical organization.

D.3 Entropy Calibration Diagnostics

Table 10 reports the lagwise uncertainty statistics. For every checkpoint and lag, signed errors are computed from the same frozen reference/generated joint before taking absolute values or averaging across lags.

Interpretation. The transfer-entropy identity is exact for every lagwise common joint. If absolute gaps are averaged across checkpoints and lags, only

$$0 \le \overline{\Delta_{\mathrm{TE}}} \le \overline{\Delta H}_{\mathrm{tr}} + \overline{\Delta H}_{\mathrm{cond}} \tag{72}$$

is guaranteed. The reported aggregate gaps are obtained after computing the signed errors at each checkpoint and lag. The marginal and conditional occupancy gaps are separate functionals and are not constrained by this temporal identity.

D.4 Entropy-Calibration Controls

This section examines the optimization contribution of reference-entropy calibration. We first scale all five γ coefficients by a_γ while retaining the profile-divergence coefficients, then compare auxiliary-gradient magnitude and direction.

Increasing a_γ from 1 to 2 brings DIV-K from 3.072 to 3.18, closer to GT 3.290, and further reduces the reported entropy gaps. FID-K, SK, and CJ nevertheless worsen. The sweep therefore identifies a quality–variation trade-off: closer scalar-gap or diversity agreement alone does not select the best task configuration. The gradient-matched comparison below separately examines aggregate auxiliary-gradient magnitude.

Matching auxiliary-gradient magnitude improves entropy-free FID-K from 5.072 to 4.91, while Full achieves 4.342. The remaining difference supports the contribution of the calibration objective beyond a comparable aggregate gradient norm.

The measured gradients are related to the profile gradient but are not collinear with it. Together with the matched-norm comparison, this supports a distinct optimization contribution from uncertainty calibration.

D.5 Safe Camera Decoder Analysis

Table 14 isolates deterministic projection from camera synthesis by comparing the raw VAE output with its projected form under the matched-output protocol.

Interpretation. The reported invalid rate changes from 0.0101 to 0.0021 and CJ from 0.174 to $0.122^{\pm 0.02}$, while FID-K changes from 4.44 to $4.342^{\pm 0.22}$ and SK from 2.72 to $2.65^{\pm 0.21}$. The larger changes occur in geometry and local derivatives. This helps distinguish the projection's role from the task improvement associated with E-KPM. The table reports matched-output mean differences.

Table 9: Concise definitions of the supervision-formulation controls and the independent tokenizer control.

Variant	Reference association	Supervision operator
T0: Paired	authored camera C_i^+ of the same music–dance sample	sample-wise camera cost $\ell_{\mathrm{cam}}(\widehat{C}_i, C_i^+)$
T1: Global set	non-paired training camera C_{j_i} from the global eligible pool, $j_i \sim \mathcal{B}_{\mathrm{global}}(i)$	same cost $\ell_{\mathrm{cam}}(\widehat{C}_i, C_{j_i})$ against the drawn reference
T2: Condition-bin set	non-paired camera from the same frozen profiling state, $j_i \sim \mathcal{B}_{Y_i}(i)$	same cost $\ell_{\mathrm{cam}}(\widehat{C}_i, C_{j_i})$ against the drawn reference
T3: Set distribution	state-conditioned reference collection $P_r(\cdot \mid y)$, without sample correspondence	set discrepancy $\sum_y \omega_y^{\mathrm{set}} \mathcal{D}_{\mathrm{set}}(\widehat{P}_g(\cdot \mid y), P_r(\cdot \mid y))$
Token-only	frozen kinetic-tokenizer targets $\{\boldsymbol{\mu}_k\}$	token-level auxiliary objective $\mathcal{R}_{\mathrm{token}}(\{\mathbf{K}_i\}, \{\boldsymbol{\mu}_k\})$

Table 10: Lagwise transfer-entropy and occupancy-entropy gaps. For every lag, $e_{\mathrm{TE},\ell} = e_{\mathrm{tr},\ell} - e_{\mathrm{cond},\ell}$ exactly; displayed gaps are lag-weighted means of absolute errors.

Variant	ΔH_M	ΔH_C	ΔH_{tr}	ΔH_{cond}	TE.g
w/o entropy calibration	0.46	0.39	0.31	0.13	0.18
EntroKine3D	**0.07**	**0.06**	**0.09**	**0.03**	**0.08**

D.6 Additional Quantitative Results

Temporal profiling scale. Table 15 reports the validation sweep of the profiling window with $H = 15$ and $I = 75$ fixed. Increasing W from 45 to 75 lowers the reported FID-K from 7.03 to 4.43, SK from 3.88 to 2.65, and CJ from 0.215 to 0.122. The evaluated combination is best at $W = 75$, which balances temporal context with local kinetic resolution.

Original short-clip protocol. Table 16 reports the original short-clip protocol, with its own GT statistics and available baselines. Table 1 separately evaluates the regrouped sequences. Among model rows, EntroKine3D has the lowest reported FID-K and DMR, with DIV-K 3.244 against GT 3.275. TemMEGA retains the lowest FID-S and DanceCamAnimator the lowest LCD. The two protocols provide complementary comparisons.

D.7 Controlled Distribution Diagnostics

The three controls below operate on different objects. Noise response changes the sampling stream under a fixed performance. Condition intervention changes the generator input before sampling. Temporal permutation changes the order of a fixed kinetic-state multiset. Keeping these operations distinct is necessary for interpreting the results.

Table 11: Entropy-calibration weight sweep. $a_\gamma = 0$ is entropy-free and $a_\gamma = 1$ is Full.

a_γ	FID-K↓	DIV-K≈GT	SK↓	Coverage↑	CJ↓	ΔH_M	TE.g↓
0.0	5.072	2.914	**2.57**	0.84	**0.118**	0.46	0.18
0.5	4.627	3.001	2.60	0.87	0.120	0.15	0.07
1.0	**4.342**	3.072	2.65	0.89	0.122	0.07	0.08
2.0	4.553	**3.182**	2.89	**0.91**	0.137	**0.04**	**0.06**

Table 12: Weight- and gradient-matched entropy control. The matched control keeps the profile objective and the total auxiliary-gradient norm within 5% of Full while setting all explicit entropy coefficients to zero.

Variant	FID-K↓	DIV-K≈GT	SK↓	Coverage↑	TE.g↓
Entropy-free	5.072	2.914	2.57	0.84	0.18
Entropy-free, gradient-matched	4.912	2.931	**2.56**	0.85	0.17
Full entropy calibration	**4.342**	**3.072**	2.65	**0.89**	**0.08**

Table 13: Gradient relation between each auxiliary functional and the profile objective, measured on validation minibatches.

Functional	Cosine with $\nabla \mathcal{L}_{\mathrm{profile}}$	Relative norm
ΔH_M	0.62	0.18
ΔH_C	0.48	0.15
ΔH_{tr}	0.31	0.11
ΔH_{cond}	0.22	0.09
Δ_{TE}	0.12	0.07

Table 14: Effect of the deterministic geometry-safe projection. Both rows share checkpoint, condition, seed, four-step sampler, and raw VAE output. Values are mean ± raw-sequence paired-bootstrap 95% interval.

Variant	FID-K	SK	Coverage	DMR	LCD	CJ	Invalid
Raw VAE decode	$4.44^{\pm 0.23}$	$2.72^{\pm 0.22}$	$0.88^{\pm 0.05}$	$0.0047^{\pm 0.0005}$	$0.229^{\pm 0.014}$	$0.174^{\pm 0.03}$	$0.0101^{\pm 0.0011}$
+ Safe projection	$\mathbf{4.342^{\pm 0.22}}$	$\mathbf{2.65^{\pm 0.21}}$	$\mathbf{0.89^{\pm 0.05}}$	$\mathbf{0.0032^{\pm 0.0004}}$	$\mathbf{0.208^{\pm 0.01}}$	$\mathbf{0.122^{\pm 0.02}}$	$\mathbf{0.0021^{\pm 0.0004}}$

Table 15: Validation sensitivity to the temporal profile window.

W	H	I	FID-K	SK	CJ
45	15	75	7.03	3.88	0.215
60	15	75	5.75	2.71	0.160
75	15	75	**4.43**	**2.65**	**0.122**
90	15	75	4.62	2.68	0.145
120	15	75	5.07	2.95	0.173

Same-condition response. Table 17 reports NS $0.34^{\pm 0.03}$ for Full versus $0.14^{\pm 0.02}$ for LCM-only. Full also has lower SK and invalid rate. NS measures within-performance variation, while SK and Invalid characterize that variation. CC.D summarizes between-performance variation under its stated cross-condition pairing rule.

Changing the conditioning input. Table 18 reports regeneration under correct, shifted, and shuffled conditions with the same checkpoint and seed protocol. Shuffling raises C.KL from 0.90 to 2.35 and FID-K from 4.342 to 5.549, while the time-shifted setting lies between them. This supports sensitivity to the correspondence with the performance. The ordering connects camera behavior to the alignment of the supplied music–dance input.

Changing order without changing occupancy. Table 19 preserves the kinetic-state multiset, so TokenKL and mean Coverage remain unchanged by construction. TransKL_5 increases from 0.15 to $3.60^{\pm 0.28}$ and local TE.g_5 from $0.09^{\pm 0.02}$ to $0.74^{\pm 0.07}$. This demonstrates the sensitivity of temporal diagnostics to ordering, and explains why marginal agreement cannot replace a temporal objective. Appendix D.8 extends the analysis with non-overlapping lags, block permutation, and past-only conditions.

D.8 Temporal Robustness Beyond Overlapping Windows

With $W_{\text{prof}} = 75$ and a five-frame separation, adjacent descriptors share $70/75 = 93.3\%$ of their frames. The corresponding TE_5 estimate is therefore local and can inherit overlap-induced autocorrelation. It is diagnostic only, not part of the final temporal loss. The reported comparison evaluates start separations $\ell \in \{5, 75, 150\}$, retains within-sequence pairs, and obtains confidence intervals by resampling raw sequences as blocks.

Transition KL increases at the two longer lags, where fewer pairs exist, but Full remains below LCM-only at each lag. Its non-overlapping TE gaps are 0.07 and 0.09, compared with 0.50 and 0.54 for LCM-only, supporting organization beyond the local overlap.

Past-only condition. To prevent Y from sharing frames with the predicted camera window, the past-only state Y_i^- is computed from the immediately preceding 75 music–dance frames. Windows without a complete past condition are excluded. Time-shifted and sequence-shuffled past states provide negative controls.

Every row in Table 21 re-profiles the *same* frozen rollout set while changing only the past-state label attached to generated windows. It therefore measures transfer dependence on Y^-; FID-K, FID-S,

Table 16: Short-sequence comparison under the original DCM clip protocol.

Method	FID_k	FID_s	DIV_k	DIV_s	DMR	LCD
GT	–	–	3.275	1.731	0.00142	–
DanceRevolution (Huang et al., 2021)	10.267	2.368	1.491	1.118	0.0062	0.154
AI Choreographer (Li et al., 2021)	5.205	0.960	1.505	1.007	0.0899	0.310
DanceCamera3D (Wang et al., 2024a)	3.749	0.280	1.631	1.326	0.0025	0.147
DanceCamAnimator (Wang et al., 2024b)	3.453	0.268	3.140	1.293	0.0022	**0.141**
TemMEGA (Xu et al., 2025)	3.237	**0.255**	1.961	1.347	0.0020	0.152
EntroKine3D	**1.862**	0.272	**3.244**	**1.509**	**0.0016**	0.150

Table 17: Noise response and cross-condition kinetic diversity on the grouped test split. Repeated rows use the same checkpoint/seed aggregation and safe projection as Table 1.

Method	NS	CC.D ≈GT	SK	Invalid
DanceCamera3D (Wang et al., 2024a)	N/A (det.)	3.85	5.52	0.006
LCM w/o E-KPM	$0.14^{\pm 0.02}$	$2.77^{\pm 0.14}$	$6.46^{\pm 0.44}$	$0.005^{\pm 1\times 10^{-3}}$
EntroKine3D	$\mathbf{0.34^{\pm 0.03}}$	$\mathbf{3.02^{\pm 0.14}}$	$\mathbf{2.65^{\pm 0.21}}$	$\mathbf{0.0021^{\pm 4\times 10^{-4}}}$

DIV, SK, Coverage, and CJ are invariant by construction and are not repeated as if a profiling label could change a generated trajectory.

Capacity-matched full-condition controls. The 16-state profiling label Y is deliberately coarser than the deployed condition $\mathbf{c} = E_{\text{cond}}(A, M)$. To separate information content from encoder capacity, every retrained null retains the unchanged E_{cond} architecture, input tensor shape, attention masks, trainable parameterization, optimizer, initialization family, batch size, 3,000-step budget, E-KPM references, and sampling protocol. *Y-only CM* maps one-hot Y through fixed, seeded, variance-matched maps into music and dance tensors of the original shape, then passes them through the unchanged encoder. *Unconditional CM* supplies a fixed zero tensor of the same shape. *Train-time derangement* preserves every full music–dance tensor but applies a fresh within-duration-and-length derangement to condition–camera assignments in each epoch, with no fixed points. *Test-time derangement* preserves the final checkpoint and sampling streams but permutes complete music–dance conditions between matched-duration raw sequences; evaluation references remain attached to the original raw sequence. None of these controls is used for model selection.

With condition-encoder capacity held fixed, the complete synchronized input yields better conditional agreement than the coarse Y label. Deranging input–camera correspondence during training or deployment degrades it further. The results support conditioning on detailed music–dance information while sharing statistical references through Y.

Jerk and multi-lag training controls. The simple smoothness baseline penalizes camera-eye and FOV third differences together with second differences of rotation increments:

$$\mathcal{L}_{\text{jerk}} = \mathbb{E}_t \|\Delta^3 \mathbf{e}_t\|_1 + \lambda_f \mathbb{E}_t |\Delta^3 f_t| + \lambda_R \mathbb{E}_t \|\Delta \boldsymbol{\omega}_t\|_1. \tag{73}$$

It can lower local jitter without matching reference state transitions. The temporal-only multi-lag row activates the final lag-$\{75, 150\}$ past-conditioned temporal branch while omitting Sinkhorn alignment, marginal occupancy, current-condition occupancy, and their entropy terms. It isolates that branch within the shared backbone.

The jerk loss obtains the lowest CJ but larger transition and TE gaps. Lag-specific objectives favor their respective scales. The temporal-only multi-lag control performs well on the two non-overlapping transition diagnostics, while Full has lower FID-K and TE.g and higher Coverage. These comparisons separate local smoothness, temporal matching, and the complete quality–variation trade-off.

Block permutation. The original token-level permutation destroys overlap continuity mechanically. We therefore operate *after* extracting the ordered (Y_i, q_i) token sequence: consecutive blocks of 15

Table 18: Condition intervention on the grouped test split. Conditions are changed before regeneration, while checkpoint, generation seed, sampler, and safe projection remain fixed. Values are mean ± raw-sequence paired-bootstrap 95% interval.

Condition	FID-K	C.KL	TE.g	CS
Shuffled	$5.549^{\pm 0.31}$	$2.35^{\pm 0.14}$	$0.61^{\pm 0.06}$	$0.44^{\pm 0.05}$
Time-shifted	$4.580^{\pm 0.24}$	$1.78^{\pm 0.11}$	$0.42^{\pm 0.05}$	$0.72^{\pm 0.06}$
Correct	$\mathbf{4.342^{\pm 0.22}}$	$\mathbf{0.90^{\pm 0.05}}$	$\mathbf{0.08^{\pm 0.02}}$	$\mathbf{1.55^{\pm 0.12}}$

Table 19: Temporal-order permutation on grouped test outputs. The generated samples and marginal state multiset are fixed; only temporal order is changed. Values are mean ± raw-sequence paired-bootstrap 95% interval.

Protocol	TokenKL	Coverage	TransKL$_5$	$\Delta H_{\mathrm{tr},5}$	TE.g$_5$
Original	$\mathbf{0.18^{\pm 0.02}}$	$\mathbf{0.89^{\pm 0.04}}$	$\mathbf{0.15^{\pm 0.03}}$	$\mathbf{0.07^{\pm 0.02}}$	$\mathbf{0.09^{\pm 0.02}}$
Shuffled	$\mathbf{0.18^{\pm 0.02}}$	$\mathbf{0.89^{\pm 0.04}}$	$3.60^{\pm 0.28}$	$0.86^{\pm 0.09}$	$0.74^{\pm 0.07}$

tokens (each token start is five frames apart, hence 75-frame-equivalent blocks) are permuted while their within-block order and token multiset remain fixed. No kinetic descriptor is re-extracted after permutation. Transitions that cross a newly created block boundary are retained and are intentionally altered.

Block permutation preserves within-block token order and most local lag-five pairs, while changing cross-boundary and lag-75 transitions. The increase in TransKL$_{75}$ from 0.31 to 2.45 shows that the longer-scale diagnostic detects ordering changes even when local token runs are retained.

D.9 Fixed-Condition Stochastic Response

The global diversity statistics in the main paper characterize variation over the complete test distribution, but do not isolate variation induced by stochastic sampling under a fixed music–dance condition. We therefore analyze fixed-condition stochastic response at the level of the original performance rather than the profiling state Y. Let $u \in \mathcal{U}$ denote a test music–dance performance, $b \in \{1,2,3\}$ denote an independently trained checkpoint, and $r \in \{1,2,3\}$ denote one of the three predeclared sampling streams. The corresponding direct camera output is denoted by $\widehat{C}_{b,u,r}$. Each sampling stream determines both the initial latent and all intermediate Gaussian perturbations used by the four-step sampler.

For each condition and checkpoint, the three sampling streams form the unordered pairs $(1,2)$, $(1,3)$, and $(2,3)$. Using the same kinetic trajectory distance d_K as the existing NS evaluator, we define

$$\begin{aligned}
\mathrm{NS}_{b,u} &= \frac{d_K(\widehat{C}_{b,u,1},\widehat{C}_{b,u,2}) + d_K(\widehat{C}_{b,u,1},\widehat{C}_{b,u,3}) + d_K(\widehat{C}_{b,u,2},\widehat{C}_{b,u,3})}{3}, \\
\mathrm{NS}_b &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \mathrm{NS}_{b,u}, \\
\overline{\mathrm{NS}} &= \frac{1}{3}\sum_{b=1}^{3} \mathrm{NS}_b, \\
s_{\mathrm{NS}} &= \sqrt{\frac{1}{2}\sum_{b=1}^{3}\left(\mathrm{NS}_b - \overline{\mathrm{NS}}\right)^2}.
\end{aligned} \tag{74}$$

The aggregation first averages the three seed pairs within each performance and then averages performances within each checkpoint. The reported standard deviation is therefore computed over checkpoint-level means rather than over individual seed pairs. This prevents sampling pairs from being treated as independent training runs.

Table 20: Multi-lag evaluation of the final EntroKine3D checkpoint and LCM without E-KPM. Values are mean ± raw-sequence paired-bootstrap 95% interval; pair ratio is relative to lag 5.

Lag	Overlap	Pair ratio	Full		LCM w/o E-KPM	
frames	%		TransKL$_\ell$	TE.g$_\ell$	TransKL$_\ell$	TE.g$_\ell$
5	93.3	1.00	$\mathbf{0.15^{\pm 0.03}}$	$\mathbf{0.09^{\pm 0.02}}$	$2.30^{\pm 0.19}$	$0.56^{\pm 0.06}$
75	0.0	0.24	$\mathbf{0.31^{\pm 0.07}}$	$\mathbf{0.07^{\pm 0.02}}$	$2.91^{\pm 0.26}$	$0.50^{\pm 0.07}$
150	0.0	0.12	$\mathbf{0.36^{\pm 0.08}}$	$\mathbf{0.09^{\pm 0.03}}$	$3.08^{\pm 0.29}$	$0.54^{\pm 0.08}$

Table 21: Past-only transfer-entropy analysis of one fixed set of final-model outputs. The frozen Train reference remains indexed by the correct unpermuted past state in every row. Values are mean ± raw-sequence paired-bootstrap 95% interval.

Past-state protocol	Pair-start conditional KL↓	TE.g↓
Correct past-only Y^-	$\mathbf{0.90^{\pm 0.05}}$	$\mathbf{0.08^{\pm 0.02}}$
Time-shifted Y^-	$1.71^{\pm 0.14}$	$0.39^{\pm 0.05}$
Sequence-shuffled Y^-	$2.22^{\pm 0.17}$	$0.55^{\pm 0.06}$

Table 25 separates the magnitude of seed-induced variation from its quality and geometric validity. A larger NS is not interpreted as uniformly preferable: stochastic response is useful only when the corresponding trajectories remain compatible with the learned kinetic support and camera-validity constraints. The comparison is therefore read jointly with FID-K, SK, DMR, CJ, and the invalid-trajectory rate. In particular, the paired-control row retains an advantage only on the point-wise dancer-fidelity metrics (DMR and LCD), whereas the entropy-free row separates profile matching from explicit uncertainty calibration.

Aggregate DIV-K includes variation across performances, temporal regions, and sampling streams. The NS comparison isolates the last component within each performance: T0 has lower response than entropy-free and Full, while Full combines greater response with the reported quality improvements. The set-level ratings in Table 4 separately assess the perceived value of these alternatives.

E Additional Cinematography and Qualitative Evaluation

The kinetic descriptor in Eq. equation 2 contains camera velocity and acceleration statistics. It does not itself assess screen composition, body visibility, absolute viewpoint, shot grammar, or music-event timing. We therefore evaluate these properties with a separate projection- and event-based protocol that never reuses the kinetic extractor or its codebook.

Qualitative comparison. Figure 4 compares camera positions and rendered framing at matched timestamps across representative baselines and Full.

For each visible dancer joint $\mathbf{p}_{t,j}$, let $(u_{t,j}, v_{t,j}, z_{t,j}) = \Pi_C(\mathbf{p}_{t,j})$ be its normalized image coordinate and camera depth. With a border margin $m = 0.05$, the per-frame dancer visibility and camera–body intersection indicators are

$$\mathrm{Vis}_t = \frac{1}{|\mathcal{J}_t|} \sum_{j \in \mathcal{J}_t} \mathbb{1}\left[z_{t,j} > 0,\ |u_{t,j}| \le 1 - m,\ |v_{t,j}| \le 1 - m\right], \tag{75}$$

$$\mathrm{Int}_t = \mathbb{1}\left[\|\mathbf{e}_t - \bar{\mathbf{p}}_t\|_2 \le r_t + 0.25\right], \tag{76}$$

where $\mathcal{J}_t$ is the valid-joint set, $\bar{\mathbf{p}}_t$ is its centroid, and $r_t = \max_{j \in \mathcal{J}_t} \|\mathbf{p}_{t,j} - \bar{\mathbf{p}}_t\|_2$ is the dancer envelope radius. DVR and DIR are the sequence averages of Vis_t and Int_t, respectively. Occlusion-invalid rate (OIR) is computed with the common rendered dancer depth map D_t: a projected joint is occluded when $z_{t,j} > D_t(u_{t,j}, v_{t,j}) + 0.05$ or lies behind the camera. OIR averages this indicator over valid joints and frames.

Let B_t be the screen-space bounding box of visible joints, and let A_t, (c_t^x, c_t^y) denote its area and center. We define a framing descriptor $\mathbf{f}_t = (\log A_t, c_t^x, c_t^y)$ and an absolute-view descriptor

Table 22: Capacity- and compute-matched full-condition controls. All rows retain the same student, condition encoder, trainable parameterization, attention masks, optimizer, 3,000-step budget, E-KPM references, and sampling protocol. CM denotes an input-only intervention retaining the complete condition-encoder architecture. Values are mean ± raw-sequence paired-bootstrap 95% interval. C.KL and TE.g use each raw sequence's unpermuted reference condition.

Condition signal	P/P_{Full}	$\text{MAC}/\text{MAC}_{\text{Full}}$	Epochs	FID-K↓	FID-S↓	C.KL↓	CS↑	TE.g↓	NS≈Full
Full $E_{\text{cond}}(A, M)$ (EntroKine3D)	1.00	1.00	3000	$\mathbf{4.342}^{\pm 0.22}$	$\mathbf{0.398}^{\pm 0.03}$	$\mathbf{0.90}^{\pm 0.05}$	$\mathbf{1.55}^{\pm 0.12}$	$\mathbf{0.08}^{\pm 0.02}$	$0.34^{\pm 0.03}$
Y-only CM input	1.00	1.00	3000	$4.96^{\pm 0.24}$	$0.535^{\pm 0.04}$	$1.37^{\pm 0.10}$	$0.80^{\pm 0.09}$	$0.19^{\pm 0.04}$	$0.24^{\pm 0.03}$
Full (A, M), train-time derangement	1.00	1.00	3000	$6.12^{\pm 0.30}$	$0.690^{\pm 0.06}$	$2.61^{\pm 0.20}$	$0.23^{\pm 0.06}$	$0.52^{\pm 0.06}$	$0.28^{\pm 0.04}$
Unconditional CM input	1.00	1.00	3000	$6.27^{\pm 0.30}$	$0.715^{\pm 0.07}$	$2.73^{\pm 0.19}$	$0.18^{\pm 0.05}$	$0.53^{\pm 0.06}$	$0.29^{\pm 0.04}$
Full (A, M), test-time derangement	1.00	1.00	frozen	$5.55^{\pm 0.28}$	$0.512^{\pm 0.05}$	$2.31^{\pm 0.17}$	$0.39^{\pm 0.07}$	$0.49^{\pm 0.06}$	$\mathbf{0.31}^{\pm 0.04}$

Table 23: Temporal-objective controls. All rows use the same backbone, sampling policy, and training budget. Values are mean ± raw-sequence paired-bootstrap 95% interval. The final row is the unique paper model; the temporal-only multi-lag row isolates its lag-$\{75, 150\}$, past-only temporal branch.

Variant	FID-K↓	Coverage↑	CJ↓	TransKL_5 ↓	TransKL_{75} ↓	TransKL_{150} ↓	Past TE.g↓
LCM only	$7.711^{\pm 0.48}$	$0.52^{\pm 0.05}$	$0.542^{\pm 0.06}$	$2.30^{\pm 0.19}$	$2.91^{\pm 0.26}$	$3.08^{\pm 0.29}$	$0.52^{\pm 0.06}$
$+\mathcal{L}_{\text{jerk}}$	$6.32^{\pm 0.31}$	$0.55^{\pm 0.04}$	$\mathbf{0.104}^{\pm 0.02}$	$1.91^{\pm 0.17}$	$2.78^{\pm 0.24}$	$2.94^{\pm 0.28}$	$0.48^{\pm 0.05}$
+ overlap transition, lag 5	$4.471^{\pm 0.19}$	$0.84^{\pm 0.05}$	$0.119^{\pm 0.02}$	$\mathbf{0.15}^{\pm 0.03}$	$1.42^{\pm 0.16}$	$1.51^{\pm 0.18}$	$0.31^{\pm 0.04}$
+ non-overlap transition, lag 75	$5.08^{\pm 0.26}$	$0.78^{\pm 0.06}$	$0.162^{\pm 0.02}$	$0.88^{\pm 0.08}$	$\mathbf{0.22}^{\pm 0.05}$	$0.91^{\pm 0.10}$	$0.24^{\pm 0.04}$
+ temporal-only multi-lag Y^- control	$4.53^{\pm 0.22}$	$0.86^{\pm 0.05}$	$0.127^{\pm 0.02}$	$0.18^{\pm 0.04}$	$0.29^{\pm 0.06}$	$\mathbf{0.35}^{\pm 0.07}$	$0.11^{\pm 0.03}$
EntroKine3D (final)	$\mathbf{4.342}^{\pm 0.22}$	$\mathbf{0.89}^{\pm 0.06}$	$0.122^{\pm 0.02}$	$\mathbf{0.15}^{\pm 0.03}$	$0.31^{\pm 0.07}$	$0.36^{\pm 0.08}$	$\mathbf{0.08}^{\pm 0.02}$

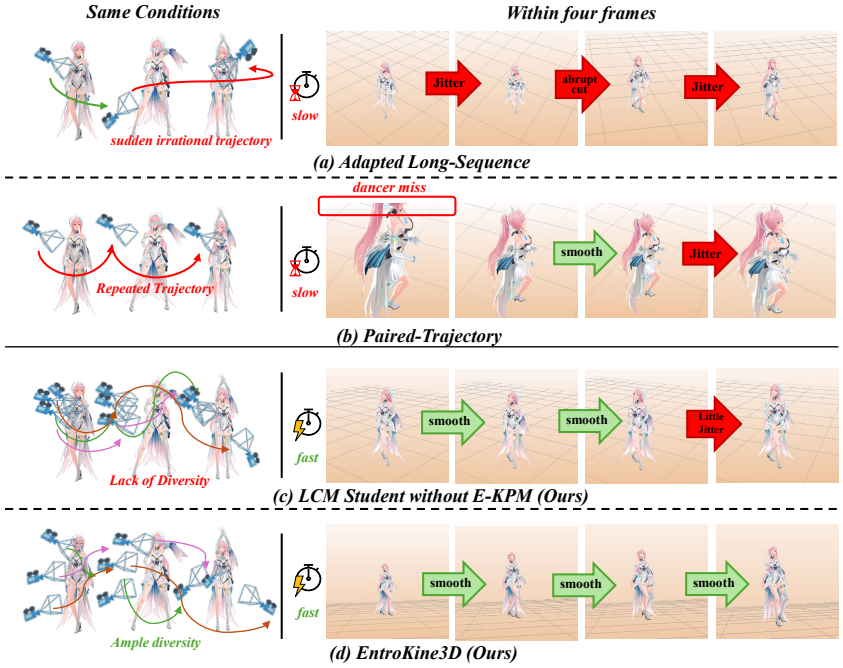


Figure 4: **Camera trajectories and rendered framing at matched timestamps.** Rows compare the same performance using representative task baselines, adapted motion backbones, LCM without E-KPM, and Full. The annotations identify local framing and motion events at the displayed timestamps.

$\mathbf{a}_t = (\log \|\mathbf{e}_t - \bar{\mathbf{p}}_t\|_2, \text{azimuth}_t, \text{elevation}_t, f_t)$. Framing composition distance and absolute-view distance are debiased Sinkhorn divergences between generated descriptors and authored-camera descriptors from the same evaluation split:

$$\text{FCD} = \mathcal{S}(P_g(\mathbf{f}), P_r(\mathbf{f})), \qquad \text{AVD} = \mathcal{S}(P_g(\mathbf{a}), P_r(\mathbf{a})). \tag{77}$$

These descriptors use projected layout and absolute camera state. Preprocessing and the shot-state quantizer are fitted on training data; held-out authored cameras supply only evaluation reference statistics, following Appendix C.

Table 24: Token-level order interventions that preserve occupancy but destroy different temporal scales. The block intervention permutes 15-token (75-frame-equivalent) blocks after token extraction. Values are mean ± raw-sequence paired-bootstrap 95% interval.

Protocol	TokenKL	Coverage	TransKL$_5$	TransKL$_{75}$
Original order	$\mathbf{0.18^{\pm 0.02}}$	$\mathbf{0.89^{\pm 0.04}}$	$\mathbf{0.15^{\pm 0.03}}$	$\mathbf{0.31^{\pm 0.07}}$
Window-token shuffle	$\mathbf{0.18^{\pm 0.02}}$	$\mathbf{0.89^{\pm 0.04}}$	$3.60^{\pm 0.28}$	$3.24^{\pm 0.31}$
15-token block permutation	$\mathbf{0.18^{\pm 0.02}}$	$\mathbf{0.89^{\pm 0.04}}$	$0.62^{\pm 0.08}$	$2.45^{\pm 0.24}$

Table 25: Fixed-condition stochastic response together with task-level quality and geometric-validity statistics.

Method	NS	FID-K ↓	SK ↓	DMR ↓	CJ ↓	Invalid ↓
DanceCamera3D (Wang et al., 2024a)	N/A (det.)	9.555	5.52	0.0044	1.067	0.006
T0: Paired	$0.08^{\pm 0.02}$	$6.381^{\pm 0.13}$	$4.73^{\pm 0.26}$	$\mathbf{0.0026^{\pm 4\times 10^{-4}}}$	$0.267^{\pm 0.021}$	$0.0030^{\pm 3\times 10^{-4}}$
LCM w/o E-KPM	$0.14^{\pm 0.02}$	$7.711^{\pm 0.48}$	$6.46^{\pm 0.44}$	$0.0045^{\pm 6\times 10^{-4}}$	$0.542^{\pm 0.06}$	$0.0057^{\pm 1\times 10^{-3}}$
EntroKine3D w/o entropy calibration	$0.26^{\pm 0.04}$	$5.072^{\pm 0.12}$	$\mathbf{2.57^{\pm 0.25}}$	$0.0037^{\pm 6\times 10^{-4}}$	$\mathbf{0.118^{\pm 0.03}}$	$0.0030^{\pm 1\times 10^{-4}}$
EntroKine3D	$\mathbf{0.34^{\pm 0.03}}$	$\mathbf{4.342^{\pm 0.22}}$	$2.65^{\pm 0.21}$	$0.0032^{\pm 4\times 10^{-4}}$	$0.122^{\pm 0.02}$	$\mathbf{0.0021^{\pm 4\times 10^{-4}}}$

For long-sequence grammar, a Train-fitted quantizer maps $(\mathbf{f}_t, \mathbf{a}_t)$ to a discrete shot state S_t. Consecutive identical states form shot segments $(\widetilde{S}_n, D_n)$ with duration D_n, so both shot transitions and cut spacing are represented. We report shot-grammar distance

$$\mathrm{SGD} = \frac{1}{2}\sum_s p_r(s) D_{\mathrm{KL}}\left[p_r(\widetilde{S}_{n+1} \mid \widetilde{S}_n = s) \| p_g(\widetilde{S}_{n+1} \mid \widetilde{S}_n = s)\right] + \frac{1}{2} W_1(P_r(D), P_g(D)). \quad (78)$$

This measures distributional shot grammar and cut duration without demanding framewise reconstruction of the single recorded shot sequence. To assess event timing, let o_t be the fixed audio onset envelope, d_t the root-relative dance-motion-energy envelope, and $g_t = \|\Delta(\mathbf{f}_t, \mathbf{a}_t)\|_1 + \mathbb{1}[S_t \neq S_{t-1}]$ the projected camera-event envelope. Event alignment is

$$\mathrm{EAS} = \frac{1}{2}\left[\max_{|\tau|\le 5} \rho(o_t, g_{t+\tau}) + \max_{|\tau|\le 5} \rho(d_t, g_{t+\tau})\right], \quad (79)$$

where ρ is Pearson correlation over valid aligned frames. EAS measures temporal correspondence with music and dance events; it is not inferred from the E-KPM condition bins.

Table 26 complements kinetic evaluation with projected framing, visibility, absolute view, shot grammar, and event correspondence. Under the common projection, Full improves the reported means relative to entropy-free across these diagnostics. Table 14 separately evaluates the effect of the projection on matched outputs.

Blinded cinematography evaluation. We additionally use a panel of 16 participants with at least two years of camera animation, virtual cinematography, editing, or 3D layout experience. The protocol contains 24 duration-balanced test performances and six direct-output methods: EntroKine3D, entropy-free, LCM without E-KPM, DanceCamera3D, DanceCamAnimator, and EDGE-Cam. Each method–condition clip receives six independent presentations, yielding 864 rater–clip presentations and 3,456 ordinal responses over composition, dancer visibility, event timing, and shot continuity. Method identity is hidden; the renderer, audio loudness, resolution, and playback settings are equalized. Repeated presentations of a raw performance to the same participant are separated by at least ten trials. Exposure is summarized by the reported presentation counts; Appendix F specifies their counting units. A crossed cumulative-link model with participant and condition effects follows Eq. equation 80; the rating guide contains only visual criteria and no kinetic metric or score.

The expert panel favors Full on the reported means for composition, visibility, event timing, and shot continuity. These judgments evaluate rendered camerawork independently of the kinetic extractor.

Table 26: Independent projection-, viewpoint-, grammar-, and event-based cinematography diagnostics on grouped DCM test sequences. FCD, OIR, DIR, AVD, and SGD are lower; DVR and EAS are higher. These measures do not use the E-KPM kinetic feature extractor or codebook.

Method	FCD↓	DVR↑	OIR↓	DIR↓	AVD↓	SGD↓	EAS↑
DanceCamera3D	0.286	0.886	0.061	0.0142	0.412	0.684	0.463
DanceCamAnimator	0.232	0.913	0.048	0.0098	0.348	0.552	0.508
EDGE-Cam	0.317	0.874	0.073	0.0171	0.471	0.801	0.481
LCM w/o E-KPM	0.251	0.901	0.055	0.0111	0.384	0.624	0.536
EntroKine3D w/o entropy calibration	0.187	0.939	0.036	0.0062	0.239	0.318	0.641
EntroKine3D	**0.149**	**0.957**	**0.024**	**0.0031**	**0.188**	**0.217**	**0.702**

Table 27: Blinded cinematography-expert evaluation on a 1–7 scale. Brackets give participant–condition cluster-bootstrap 95% intervals.

Method	Composition	Visibility	Event timing	Shot continuity
DanceCamera3D	4.78 [4.60,4.96]	4.92 [4.74,5.09]	4.69 [4.51,4.87]	4.61 [4.43,4.79]
DanceCamAnimator	5.31 [5.15,5.46]	5.52 [5.36,5.67]	5.08 [4.91,5.24]	5.19 [5.02,5.35]
EDGE-Cam	4.83 [4.65,5.00]	4.71 [4.53,4.89]	4.96 [4.79,5.14]	4.74 [4.56,4.92]
LCM w/o E-KPM	4.92 [4.75,5.10]	5.08 [4.91,5.25]	5.12 [4.95,5.29]	4.88 [4.70,5.06]
Entropy-free	5.73 [5.58,5.87]	5.96 [5.82,6.10]	5.61 [5.45,5.76]	5.68 [5.52,5.83]
EntroKine3D	**6.48 [6.34,6.61]**	**6.63 [6.50,6.75]**	**6.25 [6.10,6.39]**	**6.31 [6.17,6.45]**

F HUMAN EVALUATION

Protocol and participants. The expanded study evaluates direct outputs for 30 test conditions balanced by duration. Seven methods—Full, entropy-free, T0 Paired, LCM without E-KPM, EDGE-Cam, DanceCamera3D, and DanceCamAnimator—yield $30 \times 7 = 210$ method–condition clips. Each clip receives 12 rater presentations, producing 2,520 rater–clip presentations and 10,080 criterion-level ordinal responses because every presentation records four ratings. Forty-eight retained participants complete a balanced incomplete block: half rate 52 clips and half rate 53 clips. For stochastic systems, one stream from the three predeclared streams is assigned to each method–condition clip by a cyclic Latin-square schedule: each method uses each stream ten times overall, and the duration-group imbalance per stream is at most one clip. No participant sees a method–condition clip twice; repeated raw conditions are separated by at least 12 intervening trials and contribute at most two method outputs to one participant. Synthesis streams are frozen before inspection. Method identity is hidden, presentation order is randomized, renderer settings are shared, audio is loudness-normalized, and one replay is allowed.

Rating criteria. Each clip receives seven-point ratings (Likert, 1932) for overall quality, camera-motion naturalness, music–dance responsiveness, and dancer framing. Overall quality concerns suitability as dance camerawork; naturalness concerns smoothness and physical plausibility; responsiveness concerns the relationship between camera changes and the performance; framing concerns composition and visibility. The endpoints represent the lowest and highest agreement with each criterion.

Crossed ordinal model. For criterion q, participant p, condition u, and method m, the seven-point response is modeled with a cumulative-link mixed model,

$$\log \frac{\Pr(R_{pumq} \le \ell)}{\Pr(R_{pumq} > \ell)} = \tau_{q\ell} - \left(\beta_{mq} + \gamma_{dq} + b_{pq} + b^{(m)}_{pmq} + c_{uq} + v_{muq}\right), \quad \ell = 1, \dots, 6, \quad (80)$$

where method and duration group are fixed effects; b_{pq} and c_{uq} are crossed participant and condition random intercepts; $b^{(m)}_{pmq}$ is the preregistered participant-specific method slope; and v_{muq} is a rendered-clip intercept nested within method–condition. The prespecified fallback for a singular primary fit sets $b^{(m)}_{pmq} = 0$ while retaining the other random effects and is reported as a sensitivity analysis. Overall quality is the primary endpoint. Naturalness, responsiveness, and framing are secondary endpoints with Holm correction. We report adjusted odds ratios, cluster-bootstrap 95% intervals, corrected

Table 28: Single-output human evaluation on a 1–7 scale. Brackets give participant–condition cluster-bootstrap 95% intervals.

Method	Overall	Naturalness	Responsiveness	Framing
EntroKine3D	**6.6 [6.47,6.72]**	**6.1 [5.96,6.23]**	**6.4 [6.27,6.52]**	**6.7 [6.58,6.81]**
Entropy-free	5.8 [5.65,5.94]	5.9 [5.76,6.03]	5.6 [5.44,5.75]	6.1 [5.96,6.24]
T0 Paired	5.5 [5.34,5.65]	5.6 [5.44,5.75]	5.4 [5.23,5.56]	5.8 [5.64,5.95]
EDGE-Cam (Tseng et al., 2023)	5.2 [5.03,5.36]	5.1 [4.93,5.26]	5.0 [4.82,5.17]	4.90 [4.72,5.07]
DanceCamera3D (Wang et al., 2024a)	4.9 [4.72,5.07]	4.3 [4.12,4.48]	5.1 [4.93,5.26]	4.8 [4.62,4.97]
DanceCamAnimator (Wang et al., 2024b)	4.7 [4.52,4.88]	4.9 [4.72,5.07]	4.7 [4.52,4.88]	4.6 [4.42,4.78]
LCM w/o E-KPM	4.3 [4.12,4.48]	4.6 [4.42,4.78]	4.2 [4.02,4.38]	4.2 [4.02,4.38]

Table 29: Adjusted ordinal odds ratios for EntroKine3D against the two mechanism controls and strongest adapted comparator. Brackets give 95% intervals; p_{H} is Holm-corrected.

Criterion	Comparator	OR [95% CI]	p_{H}
Overall	Entropy-free	2.05 [1.65,2.55]	< 0.001
Overall	T0 Paired	2.78 [2.21,3.50]	< 0.001
Overall	EDGE-Cam	3.60 [2.65,4.91]	< 0.001
Naturalness	Entropy-free	1.29 [1.06,1.58]	0.018
Naturalness	T0 Paired	1.64 [1.31,2.05]	< 0.001
Naturalness	EDGE-Cam	2.48 [1.88,3.29]	< 0.001
Responsiveness	Entropy-free	2.16 [1.72,2.72]	< 0.001
Responsiveness	T0 Paired	2.75 [2.16,3.50]	< 0.001
Responsiveness	EDGE-Cam	3.32 [2.45,4.51]	< 0.001
Framing	Entropy-free	1.84 [1.45,2.32]	< 0.001
Framing	T0 Paired	2.42 [1.91,3.07]	< 0.001
Framing	EDGE-Cam	4.57 [3.32,6.30]	< 0.001

p-values, ordinal Krippendorff's α, and ICC. Participants and conditions are cluster-resampled for robustness analysis.

Same-condition set evaluation. An additional task uses 18 disjoint test conditions and four stochastic systems: Full, entropy-free, T0 Paired, and LCM without E-KPM. Each displayed set contains three unfiltered direct sampling streams for the same music–dance performance. The 72 method–condition sets receive six rater presentations each, yielding 432 rater–set presentations and 1,296 criterion-level ordinal responses; every participant rates nine sets. Raters separately judge naturalness of all three samples, condition consistency, and effective diversity, defined as visibly different interpretations without invalid/implausible members.

Table 4 (Section 4.4) reports the resulting criterion means with cluster-bootstrap 95% intervals for all four systems.

Agreement and quality control. Agreement is summarized by ordinal Krippendorff's α and ICC. The respective (α, ICC) pairs are $(0.61, 0.73)$ for overall quality, $(0.56, 0.69)$ for naturalness, $(0.54, 0.67)$ for responsiveness, and $(0.64, 0.76)$ for framing. The set-level values are $0.60/0.72$, $0.58/0.70$, and $0.66/0.78$ for naturalness, condition consistency, and effective diversity. The preregistered exclusion rules cover failed attention checks, less than 80% viewing time, duplicate submissions, and impossible response times. Participant and clip identifiers are anonymized; recruitment source, demographics, compensation, exclusions, and institutional approval or exemption are released with the study manifest.

Presentation units. In a human study, a presentation is one rater viewing one clip or one three-sample set; its criterion ratings are correlated responses within that presentation. The single-output study has $30 \times 7 \times 12 = 2{,}520$ presentations and four ratings per presentation. The set study has $18 \times 4 \times 6 = 432$ presentations and three ratings per presentation. The cinematography panel has $24 \times 6 \times 6 = 864$ presentations and four ratings per presentation, yielding $3{,}456$ ordinal responses.

G Reproducibility, Runtime, and Implementation Details

Configuration selection. The profiling-window grid is $W_{\text{prof}} \in \{45, 60, 75, 90, 120\}$. The sampling grid is $\eta \in \{0, 0.25, 0.5, 0.75, 1\}$, and the entropy grid is $a_\gamma \in \{0, 0.5, 1, 2\}$. Each validation sweep keeps the remaining settings fixed. Ties within 0.05 FID-K are resolved by SK and then TE.g. The selected configuration is $W_{\text{prof}} = 75$, $\eta = 0.5$, and $a_\gamma = 1$. The same validation protocol selects EDGE-Cam as the adapted human-study comparator by its lowest FID-K among the three adapted backbones.