MOTIONSCM: STRUCTURAL CAUSAL MODELING FOR INTERVENTIONAL HUMAN MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing human motion control interfaces specify semantic or kinematic properties without exposing cross-joint coordination as an independently editable mechanism. This makes it difficult to isolate an action's dependence on a particular coordination rule and trace how changes to that rule propagate across the body. We introduce **MotionSCM**, a motion generator for studying these dependencies through mechanism interventions. Its time-varying structural causal model uses explicit structural assignments to generate each motion channel. Under stated causal assumptions, post-training edits replace selected assignments while holding generator inputs and unedited mechanisms fixed. Per-channel gain constraints preserve rollout stability within the admissible operator set. On professional MoCap data from six industrial and craft processes, the learned structure varies across actions and individual motions. Interventions on the same rule produce distinct action-dependent cross-joint responses, with lagged recurrence sustaining the transmitted changes. MotionSCM retains competitive generation quality on three action-conditioned benchmarks. Structural ablations further show that improved generation fit need not preserve intervention behaviour. These results establish coordination as an explicit object of causal analysis within the motion generator. Code is available at anonymous.4open.science/r/MotionSCM-736C.

1 INTRODUCTION

Recent human motion generators support precise control through text, action labels, target trajectories, and joint-level constraints (Khani et al., 2025; Tevet et al., 2023; Zhang et al., 2023a; Xie et al., 2024). Motion-language models combine generation with understanding and refinement (Li et al., 2026; Wang et al., 2026b; Li et al., 2025a). Fine-grained editors extend these capabilities to local changes in generated motion (Pinyoanuntapong et al., 2024; Athanasiou et al., 2024; Zhang et al., 2023b; Xie et al., 2024; Chen et al., 2023; Hong et al., 2025). These advances provide increasingly detailed control over the semantic and kinematic properties of motion.

Human motion involves coordinated movements across body parts, a dependency explicitly modeled by recent part-based generators (Zou et al., 2024) Even with a fixed task objective, simulated characters can adapt their motion to weakened local actuation (Bae et al., 2025). During repetitive sawing, fatigue-related reductions in elbow movement can be accompanied by increased movement at other joints while the tool trajectory remains largely unchanged (Côté et al., 2002). Understanding such adaptation involves more than reproducing the resulting trajectories, it requires examining how individual coordination relations contribute to the movement. Existing editing interfaces do not expose these relations as independently editable mechanisms. This leaves a gap between generating variations of an action and testing its dependence on a particular coordination relation. In a generated weaving motion, for example, weakening a forearm coordination rule would test its influence on movement elsewhere in the body.

We therefore ask: how strongly does a generated action depend on a particular coordination rule? We propose **MotionSCM**, a motion generator that represents learned coordination mechanisms through a Structural Causal Model (SCM). Selected rules can be modified after training while generator inputs and unedited mechanisms remain fixed. Under explicit causal assumptions, the resulting cross-joint response quantifies dependence on the edited rule within the generator (Pearl, 2009; Peters et al., 2017). We establish conditions that preserve rollout stability for admissible interventions.

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.

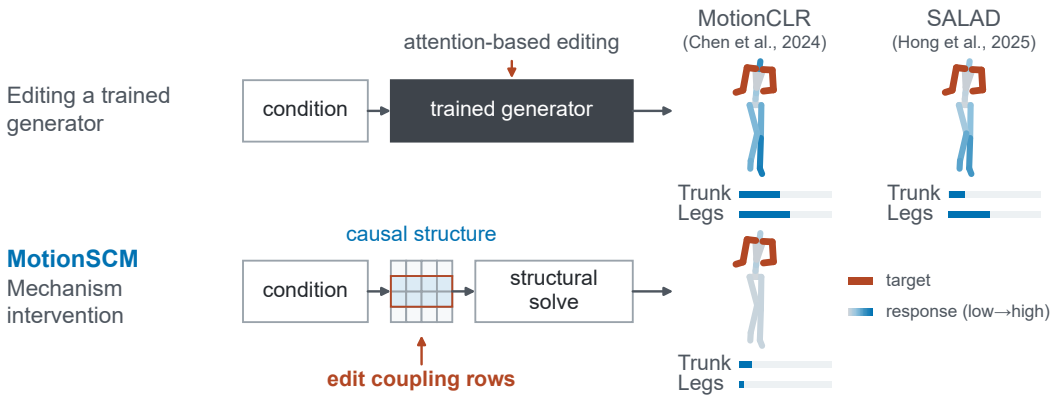

Figure 1: Attention-based editing and mechanism intervention at matched relative arm effects on walking motions. Red marks the edited arms, while blue shows the mean joint response normalized by the arm response on a shared reference pose. Bars report the corresponding root mean square (RMS) response for the trunk/head and legs. Results are averaged over three sampling seeds per method.

This allows coordination to be studied through its effects on generated motion within the same model that produces it. We evaluate this formulation under action conditioning, with limited in-domain data for each activity. Figure 1 contrasts this mechanism-level control with attention-based editing.

Our contributions are: (1) **Coordination as a causal mechanism.** We formulate cross-joint coordination as an intervenable mechanism within a motion generator. Under explicit causal assumptions, cross-joint responses quantify an action's dependence on an edited coordination rule. (2) **Stable interventional motion generation.** A latent-conditioned structural causal decoder supports post-training mechanism and value interventions. We derive sufficient per-channel gain constraints that bound the represented motion and intervention-induced deviations. (3) **Action-specific causal dependence.** We show that generated actions depend differently on the same coordination rule, with additional structural variation within each action. Path ablations establish the contribution of lagged recurrence to propagation. Structural ablations distinguish improved distributional fit from stable, recursive mechanism control. MotionSCM also retains competitive generation quality and yields lower mean off-target responses in the matched editor comparison.

2 Related Work

Motion generation and coordination modeling. Human motion generators use VAEs, discrete tokens, diffusion, or flow matching (Petrovich et al., 2021; Zhang et al., 2023a; Guo et al., 2024; Tevet et al., 2023; Zhang et al., 2024; Zhou et al., 2024; Khani et al., 2025). Editing operates on conditions or learned representations (Guo et al., 2025; Athanasiou et al., 2024; Li et al., 2025b; Jiang et al., 2025b; Zhang et al., 2023b; Shafir et al., 2024), spatial or part-level constraints (Xie et al., 2024; Pinyoanuntapong et al., 2025; Shen et al., 2026), attention (Hong et al., 2025; Chen et al., 2024), or latent guidance (Uchida et al., 2025). Robotic approaches generate task-conditioned flows on motion manifolds (Lee et al., 2025) or controllers in parameter space (Rocca et al., 2025). Part-based approaches coordinate separate motion generators (Zou et al., 2024) or learn adaptable motion priors for individual body parts in physics-based character control (Bae et al., 2025). For explicit modeling of movement organization, the Gesture Operational Model describes dynamic dependencies between body parts through a state-space formulation (Manitsaris et al., 2020; Olivas-Padilla et al., 2024). MotionSCM extends this formulation with causality.

Causal generative models and intervention semantics. SCMs define interventions by replacing selected structural assignments (Pearl, 2009; Peters et al., 2017; Eberhardt & Scheines, 2007; Correa & Bareinboim, 2020). Structural time series models distinguish contemporaneous and lagged dependencies (Sims, 1980; Lütkepohl et al., 2005; Pamfil et al., 2020; Gong et al., 2023). Deep causal generators support interventional or counterfactual sampling (Kocaoglu et al., 2018; Pawlowski et al.,

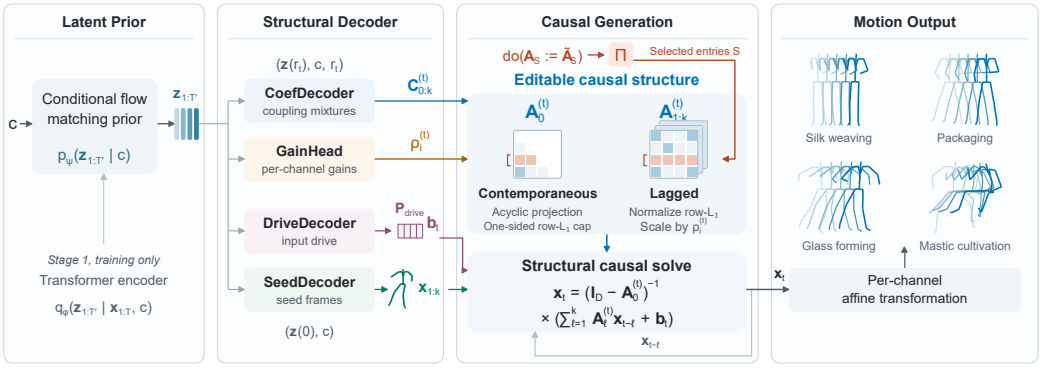

Figure 2: MotionSCM architecture. A conditional flow matching prior samples the latent trajectory. The structural decoder produces motion-specific operators, drive, and seed frames. Mechanism edits replace selected operator entries before the stability projection Π. A recursive structural solve generates normalized motion, which a fixed per-channel affine map converts to physical units.

2020; Javaloy et al., 2023; Chao et al., 2024), while causal representation learning studies latent factors (Yang et al., 2021; Brehmer et al., 2022; Lippe et al., 2023). CausalTime generates graph-labelled sequences for discovery benchmarks (Cheng et al., 2024); DoFlow and CaTSG address interventional and counterfactual prediction and generation (Wu et al., 2026; Xia et al., 2025).

In human motion, TriC-Motion uses feature disentanglement to suppress motion-irrelevant information (Cao et al., 2026). AttrMoGen separates action semantics from human attributes through an SCM-inspired information bottleneck and counterfactual attribute substitution (Wang et al., 2026a). The causal interpretation of feature operations depends on assumptions connecting them to interventions on the modeled mechanisms. Temporal causal masking has a different role, restricting access to future frames (Fu et al., 2026; Jiang et al., 2025a; Yu et al., 2026). MotionSCM intervenes on motion channel structural assignments, with explicit assumptions supporting their causal interpretation.

Stability of learned dynamics under intervention. Prior work learns stable task sequences with linear parameter-varying systems (Medina & Billard, 2017) and establishes recurrent network approximation results for incrementally stable dynamics (Hanson & Raginsky, 2020). For mechanism editing, the guarantee must hold for the executed post-intervention dynamics. MotionSCM's per-channel gain constraints define an admissible operator set and bound both rollouts and intervention-induced deviations.

3 Method

3.1 Generation Pipeline and Structural Decoder

For a conditioning signal c, let $\mathbf{x}_{1:T}$ denote the generated motion, with normalized state $\mathbf{x}_t \in \mathbb{R}^D$. Each channel represents one joint coordinate or rotation component. Figure 2 summarizes the generation pipeline.

Latent representation and prior training. Training proceeds in two stages. Stage 1 uses a Transformer encoder $q_\phi(\mathbf{z}_{1:T'} \mid \mathbf{x}_{1:T}, c)$ to infer a latent sequence $\mathbf{z}_{1:T'}$, where $\mathbf{z}_t \in \mathbb{R}^{d_z}$. The structural decoder is trained jointly by reconstructing the observed motion, with a KL regularizer toward a unit Gaussian prior.

Stage 2 fits a conditional flow matching prior $p_\psi(\mathbf{z}_{1:T'} \mid c)$ to the inferred latent trajectories. At generation time, a sampled trajectory $\mathbf{z}_{1:T'} \sim p_\psi(\cdot \mid c)$ parameterizes the structural decoder.

Latent-conditioned causal structure. At rollout time t, the sampled latent trajectory is interpolated at relative time $r_t \in [0, 1]$. The DriveDecoder maps $(\mathbf{z}(r_t), c, r_t)$ to the input drive

$\mathbf{b}_t = \mathbf{P}_{\text{drive}}\,\text{DriveDecoder}(\mathbf{z}(r_t), c, r_t)$, where $\mathbf{P}_{\text{drive}} \in \mathbb{R}^{D\times R}$ maps the R-dimensional decoder output to the D motion channels.

The same input is processed by a shared CoefDecoder. Its CoefHead produces mixture weights $\mathbf{w}(t)$, while the GainHead produces gain logits $\mathbf{g}_t$. The weights combine M shared basis matrices as $\mathbf{C}_\ell^{(t)} = \sum_{m=1}^{M} w_m(t)\mathbf{B}_{m,\ell}$, with $\mathbf{B}_{m,\ell} \in \mathbb{R}^{D\times D}$.

For each target channel i, the lagged mixtures $\mathbf{C}_{1:k}^{(t)}[i,:]$ define a coupling direction $\mathbf{d}_{1:k,i}^{(t)}$, normalized jointly over lags and source channels. Its strength is given by $\rho_i^{(t)} = q\,\tau_{\text{lag}}\,\sigma(g_{t,i})$, where σ is the logistic function, q is the operational reserve, and τ_{lag} is the lagged coupling budget. The lagged operators are then $\mathbf{A}_\ell^{(t)}[i,:] = \rho_i^{(t)}\mathbf{d}_{\ell,i}^{(t)}$.

The lag-0 mixture $\mathbf{C}_0^{(t)}$ produces the contemporaneous operator $\mathbf{A}_0^{(t)}$ after the acyclicity projection and row-L_1 cap described below. A separate SeedDecoder initializes $\mathbf{x}_{1:k} = \text{SeedDecoder}(\mathbf{z}(0), c)$. All structural quantities depend on the latent trajectory, conditioning signal, and time, but not on the generated state.

Structural causal solve. Given the structural operators and input drive above, each motion state is generated by solving

$$\mathbf{x}_t = \mathbf{A}_0^{(t)}\mathbf{x}_t + \sum_{\ell=1}^{k}\mathbf{A}_\ell^{(t)}\mathbf{x}_{t-\ell} + \mathbf{b}_t = (\mathbf{I}_D - \mathbf{A}_0^{(t)})^{-1}\left(\sum_{\ell=1}^{k}\mathbf{A}_\ell^{(t)}\mathbf{x}_{t-\ell} + \mathbf{b}_t\right). \tag{1}$$

Row i of Eq. (1) defines channel i's structural assignment, $x_{t,i} = f_{t,i}(\mathbf{x}_{\text{pa}(i)})$, with contemporaneous and lagged parents given by its nonzero coefficients. The operators $\{\mathbf{A}_0^{(t)}, \ldots, \mathbf{A}_k^{(t)}\}$ specify the coordination rule; $\mathbf{b}_t$ supplies the evolving motion drive.

Replacing a row constitutes a mechanism intervention on the corresponding structural assignment (Eberhardt & Scheines, 2007; Correa & Bareinboim, 2020). We measure its effect on the unedited channels while holding $\mathbf{z}$, c, the seed frames, and $\mathbf{b}_t$ fixed. This effect defines the cross-joint response, whose magnitude measures an action's dependence on the edited coordination rule.

The contemporaneous operator represents interactions below the sampling interval after temporal aggregation (Hyttinen et al., 2016). We threshold the time-averaged coefficient magnitudes, then order channels by outgoing strength minus incoming strength (Algorithm 1). Retaining edges that follow this per-sequence order makes $\mathbf{A}_0^{(t)}$ strictly triangular and $\mathbf{I}_D - \mathbf{A}_0^{(t)}$ invertible. Its rows are additionally L_1-capped for rollout stability.

For rollout analysis, we define $\mathbf{F}_\ell^{(t)} = (\mathbf{I}_D - \mathbf{A}_0^{(t)})^{-1}\mathbf{A}_\ell^{(t)}$ and $\mathbf{u}_t = (\mathbf{I}_D - \mathbf{A}_0^{(t)})^{-1}\mathbf{b}_t$. Then $\mathbf{x}_t = \sum_{\ell=1}^{k}\mathbf{F}_\ell^{(t)}\mathbf{x}_{t-\ell} + \mathbf{u}_t$. The guarantee uses these solved quantities; interventions modify the structural assignments. A fixed per-channel affine transformation maps the normalized state $\mathbf{x}_t$ to physical motion units.

Causal semantics and assumptions. The decoder defines a recursive SCM within the generator. Its exogenous state comprises the conditioning signal, latent trajectory, seed frames, and drive. This state includes all common causes (causal sufficiency) and retains its distribution under intervention (exogeneity). Interventions replace selected assignments and retain the others (modularity). With the per-sequence acyclic ordering, these assumptions give mechanism and value interventions their causal semantics. Counterfactual motions solve the modified assignments with the same exogenous realization (Pearl, 2009; Peters et al., 2017). Formal conditions are in Appendix A.1.1; rollout boundedness is established separately.

3.2 Per-channel Coupling Gain and Rollout Stability

The solved per-channel gain bounds propagation through the preceding k states. With $\|\mathbf{M}\|_\infty = \max_i \sum_j |M_{ij}|$, define the solved per-channel gain at frame t as $G_t = \max_i \sum_{\ell=1}^{k}\big\|\mathbf{F}_\ell^{(t)}[i,:]\big\|_1 = \big\|\sum_{\ell=1}^{k}|\mathbf{F}_\ell^{(t)}|\big\|_\infty$. For two rollouts governed by the same solved operators and drive, their deviation $\boldsymbol{\delta}_t \in \mathbb{R}^D$ satisfies $\|\boldsymbol{\delta}_t\|_\infty \le G_t \max_{1\le\ell\le k}\|\boldsymbol{\delta}_{t-\ell}\|_\infty$.

For time-varying operators, we use a uniform rollout bound (Vidyasagar, 2002), $\bar{g} := \sup_{1 \le t \le T} G_t$. The gain parameterization caps each target channel's lagged coupling as $\sum_{\ell=1}^{k} \|\mathbf{A}_\ell^{(t)}[i,:]\|_1 \le \rho_i^{(t)} < q\,\tau_{\text{lag}} \le \tau_{\text{lag}}$, while the contemporaneous operator is constrained separately by $\|\mathbf{A}_0^{(t)}\|_\infty \le \tau_{\text{con}}$. Since $\|(\mathbf{I} - \mathbf{A}_0^{(t)})^{-1}\|_\infty \le (1 - \tau_{\text{con}})^{-1}$, these budgets give the uniform ceiling $\bar{g} \le g_\star := \tau_{\text{lag}}/(1 - \tau_{\text{con}}) < 1$.

Proposition 1 (Uniform rollout stability)**.** *Assume that, for every frame $t \le T$, $\|\mathbf{A}_0^{(t)}\|_\infty \le \tau_{\text{con}} < 1$ and $\max_i \sum_{\ell=1}^{k} \left\|\mathbf{A}_\ell^{(t)}[i,:]\right\|_1 \le \tau_{\text{lag}}$, with $\tau_{\text{lag}} + \tau_{\text{con}} < 1$. Then $\bar{g} \le g_\star < 1$. Consider any two rollouts governed by the same, possibly edited, solved operators and drive. Their deviation satisfies $\|\boldsymbol{\delta}_{s+h}\|_\infty \le \bar{g}^{\lfloor h/k \rfloor} \max_{0 \le \ell < k} \|\boldsymbol{\delta}_{s+\ell}\|_\infty$ for integers $1 \le s \le T - k + 1$ and $0 \le h \le T - s$.*

Sketch. Let M_j denote the maximum deviation over the j-th k-frame window. The uniform gain bound gives $M_{j+1} \le \bar{g} M_j$. Iterating this inequality yields the stated decay. Appendix A.1.5 gives the full proof. □

Corollary 1 (Bounded state and intervention effect)**.** *Assume the conditions of Proposition 1 for both the original and executed post-edit operators, and fix a realization with bounded seed frames and solved drive. Then (i) the represented motion remains bounded, $\|\mathbf{x}_t\|_\infty \le X$; and (ii) the effect of a mechanism intervention is bounded by its induced operator and drive changes, $\max_{t \le T} \|\boldsymbol{\delta}_t\|_\infty \le (\Delta_F X + \Delta_u)/(1 - \widetilde{g})$, where Δ_F and Δ_u measure the solved operator and drive perturbations and $\widetilde{g} \le g_\star < 1$ is the post-edit uniform gain.*

The guarantees are stated on the normalized motion state and transfer to physical units through each channel's affine decode. Since AIMove's root heading is recovered by cumulative sum, its bound accumulates the per-frame bounds over the rollout.

3.3 The Intervention Interface

Mechanism intervention. Let $S \subseteq \{(t, \ell, i, j)\}$ index selected entries of the structural operators, where $\ell = 0$ denotes contemporaneous coupling and $\ell \in \{1, \dots, k\}$ denotes lagged coupling. The intervention $\text{do}(\mathbf{A}_S := \widetilde{\mathbf{A}}_S)$ replaces those coefficients and leaves the structural assignment of every unedited channel, and the raw drive $\mathbf{b}_t$, unchanged. Contemporaneous entries stay within the ordering $\prec_u$, so $\mathbf{A}_0^{(t)}$ remains strictly triangular. Its effect on the motion is bounded by Corollary 1.

Admissible operator set. The executed operators must satisfy the stability constraints of Section 3.2. Before rollout, the edited operators are projected onto the admissible operator set, $\mathbf{A}^{\text{exec}} = \Pi(\widetilde{\mathbf{A}})$, by rescaling any coupling row that exceeds its budget. Attenuating edits are unchanged; the reserve $q < 1$ permits amplification up to $1/q$ without rescaling. We use $k = 5$, $\tau_{\text{lag}} = 0.64$, $\tau_{\text{con}} = 0.35$, and $q = 2/3$. Ordinary generation and attenuation retain the operational caps, giving $\bar{g} \le q\tau_{\text{lag}}/(1 - q\tau_{\text{con}}) \approx 0.557$.

Value intervention. A value intervention $\text{do}(\mathbf{x}_{s,J} := \mathbf{v}_J)$ leaves the coupling operators unchanged and therefore does not require the projection Π. The imposed values act through the original structural dynamics, with effect decay bounded by Proposition 1. Evaluated AIMove responses remain below this bound and decay faster than its worst-case rate (Appendix A.3.1).

4 Experiments

We evaluate task-specific coordination under limited in-domain data. AIMove provides professional activities for structure and intervention analysis. HumanAct12 and CMU MoCap extend the generation comparison to different action sets and motion representations. Models use fixed class-stratified 85/15 training/validation splits. Generation metrics use the full real corpus as their reference.

4.1 The Learned Causal Structure Is Action- and Motion-Specific

Data. AIMove records professional operators across six industrial and craft processes (Olivas-Padilla et al., 2023; 2024). Its 1002 sequences at 90 Hz span 25 action classes, with a median of 18

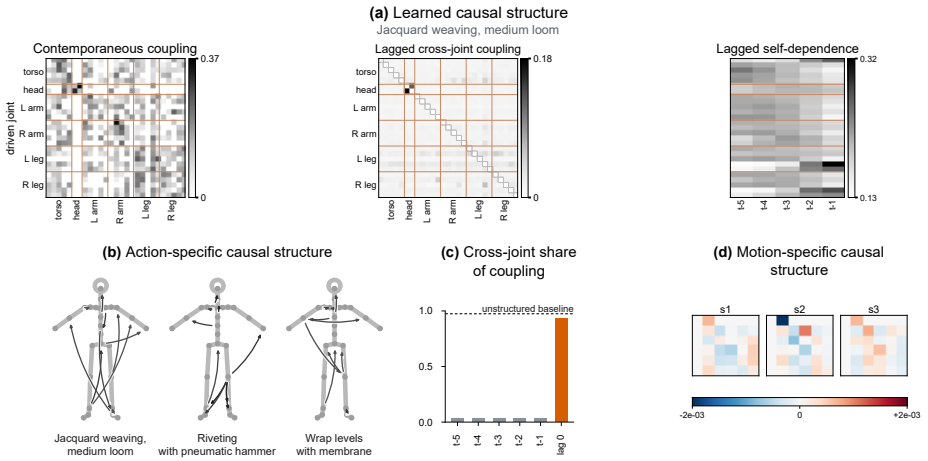

Figure 3: Learned causal structure. *(a)* Contemporaneous and lagged coupling for Jacquard weaving; *(b)* action-specific deviations of the causal structure; *(c)* cross-joint coupling share by lag; *(d)* motion-specific deviations of the causal structure. Panels *(a)* to *(c)* average over frames, sampled latents and three training seeds; *(d)* shows three sampled motions of one model against their mean.

sequences per class. We represent 27 joints by 81 rotation channels; root translation is omitted and Hips yaw is encoded by increments.

Across the lagged operators, 94% of the coefficient weight connects each channel to its own past. Direct cross-joint coupling is concentrated in the contemporaneous operator (Figure 3a,c). This allocation describes direct dependencies, not the total response after recurrent propagation.

The learned causal structure varies systematically across actions (Figure 3b). Relative to the corpus-average structure, Jacquard weaving shows increased lower-body coupling and links the left arm to the treadle leg. Riveting emphasizes coordination between the trunk and tool arm, while pallet wrapping couples arm motion with the stepping legs. Within the learned generator, these dependence patterns complement motion trajectories as a description of action organization.

Coordination structure also varies among motions of the same action. Samples of Jacquard weaving exhibit distinct deviations from their mean coupling pattern (Figure 3d). The learned structure therefore captures variation in how individual motions realize the same action.

4.2 Editing Coordination Rules and the Cross-Joint Response

We attenuate the right forearm Z rotation row by 0.5 across contemporaneous and lagged operators, retaining the latent trajectory, seed frames, and drive. We use 32 prior draws per action, 180-frame rollouts, and three training seeds. Cross-joint response is the mean absolute relative change in range of motion (ROM) across unedited channels, measured over frames 6–25. For this fixed assignment and attenuation strength, response variation across actions exceeds variation across training seeds (Figures 4 and 5). Action recognizer agreement remains 0.74 ± 0.02 for both base and edited rollouts. Leg and trunk interventions are illustrated in Appendix A.3.1.

Propagation through preserved dynamics. To separate transmission from target-amplitude changes, we fix the target's edited-minus-base trajectory and contemporaneous operators, then resolve the unedited-channel dynamics. Across 108 edits from four actions, removing lagged recurrence within unedited channels leaves 56.2% of their full-dynamics RMS response; removing only direct lagged target inputs leaves 92.7% (Table 1). Thus recurrence sustains a transmitted perturbation even though most lagged coefficient mass is within-channel.

Editor comparison. We compare mechanism interventions with the attention-based editors MotionCLR and SALAD (Chen et al., 2024; Hong et al., 2025) on walking, running, and boxing, with five sources per method and action. They use HumanML3D checkpoints; MotionSCM uses its Hu-

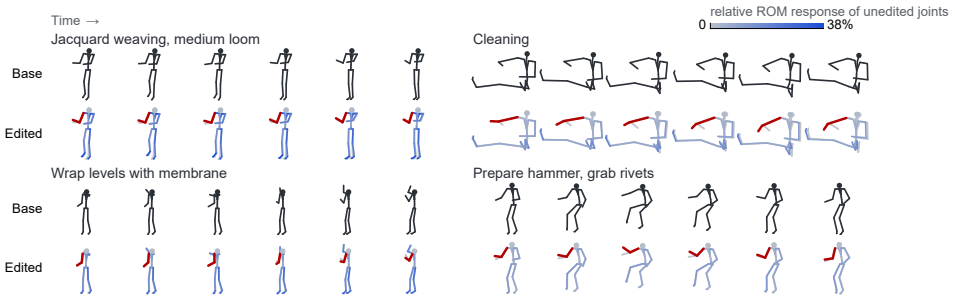

Figure 4: The same forearm row intervention across four AIMove actions. Each block pairs base (top) and edited (bottom) motion from the same latent, with the base pose ghosted. Red marks the edited arm; other bones show relative change in range of motion (ROM) over frames 6–25 on a shared percentage scale.

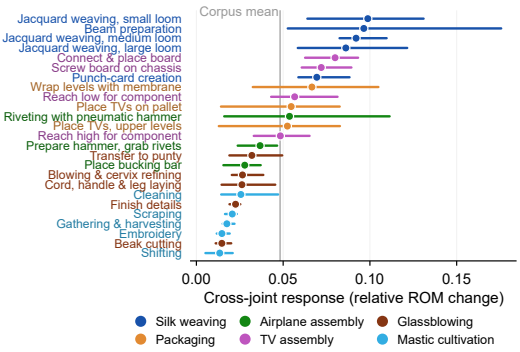

Figure 5: Per-action dependence on the edited forearm rule. Dots show mean cross-joint ROM response and bars span three training seeds. Colours denote six professional processes.

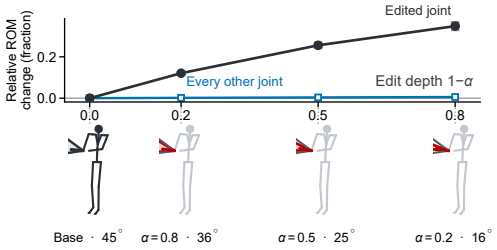

Figure 6: Fatigue-inspired forearm attenuation. Black shows one minus the edited-channel ROM ratio; blue shows unedited-channel response. Means span 25 actions and three training seeds; poses illustrate one motion.

manAct12 checkpoint. We hold each source's latent input or sampling noise fixed and match both arms' RMS displacement to 9–11% of their RMS variation before editing. On a common pelvis-relative skeleton, the lower off-target-to-arm ratios show that a comparable arm effect accompanies less change elsewhere (Table 2). SALAD boxing does not reach the matching interval. This comparison measures how edits spread across the body in each method's own motions.

Fatigue-inspired mechanism attenuation. Motivated by repetitive-work adaptation (Côté et al., 2002), we attenuate fluctuations in the forearm coupling input using $\mathrm{do}(f_{\theta,t,i} := \bar{c}_i + \alpha(c_{t,i} - \bar{c}_i) + b_{t,i})$, where $c_{t,i}$ is the coupling input, $\bar{c}_i$ its base-rollout reference mean, and $\alpha \in \{0.8, 0.5, 0.2\}$ is the mechanism attenuation parameter. The raw drive is unchanged, and target ROM is measured over frames $5k + 1$–180. Across 25 actions and three training seeds, stronger attenuation reduces mean edited-channel ROM while early unedited responses remain small (Figure 6).

4.3 Generation Quality

Data. HumanAct12 contains 1191 sequences from 12 actions and uses 72 Cartesian channels. CMU MoCap contains 1088 sequences from eight actions and uses 60 Cartesian channels (Guo et al., 2020).

We evaluate whether structural mechanisms retain competitive action-conditioned generation quality. ACTOR, MDM, MLD, and EMDM (Petrovich et al., 2021; Tevet et al., 2023; Chen et al., 2023; Zhou et al., 2024) cover variational and diffusion-based generation. All methods share motion representations and training splits. Baselines are adapted to each dataset's channels, omitting losses

Table 1: Fixed-target propagation. Target difference and contemporaneous operators are fixed. Unedited-channel RMS is relative to full dynamics (1.0); mean ± population SD over three training seeds.

Dynamics on U	Relative RMS
No lagged inputs	0.519 ± 0.030
No lagged recurrence	0.562 ± 0.012
No target-to-U lags	0.927 ± 0.064

Table 2: Matched-effect editing. Off-target-to-arm RMS ratios (mean ± population SD). Each method-action pair has five matched sources except SALAD boxing (0/5), shown as a dash.

Method	Walking	Running	Boxing
MotionSCM	$\mathbf{0.118} \pm 0.007$	$\mathbf{0.122} \pm 0.005$	$\mathbf{0.111} \pm 0.008$
MotionCLR	0.707 ± 0.135	0.830 ± 0.263	0.814 ± 0.225
SALAD	0.706 ± 0.672	0.379 ± 0.067	—

Table 3: Action-conditioned generation. Real entries are mean ± 95% confidence interval from 20 resamples. Methods report mean ± population standard deviation over three training seeds. Bold/underlining mark best/second-best; → means proximity to Real.

Dataset	Metric	Real	ACTOR	MDM	MLD	EMDM	MotionSCM
HumanAct12	FID ↓	0.36 ± 0.04	1.24 ± 0.07	$\underline{0.58} \pm 0.03$	0.75 ± 0.02	2.77 ± 0.14	$\mathbf{0.47} \pm 0.01$
	Acc. ↑	0.98 ± 0.00	0.83 ± 0.01	0.90 ± 0.00	$\mathbf{0.96} \pm 0.00$	0.28 ± 0.01	$\underline{0.93} \pm 0.01$
	Div. →	6.85 ± 0.05	6.68 ± 0.04	$\underline{6.86} \pm 0.06$	$\mathbf{6.85} \pm 0.01$	6.65 ± 0.05	6.79 ± 0.04
	MMod. →	2.64 ± 0.04	3.50 ± 0.03	3.13 ± 0.02	$\mathbf{2.40} \pm 0.01$	5.38 ± 0.03	$\underline{2.94} \pm 0.06$
AIMove	FID ↓	0.37 ± 0.05	$\underline{0.55} \pm 0.02$	0.56 ± 0.07	0.63 ± 0.02	2.96 ± 0.20	$\mathbf{0.48} \pm 0.08$
	Acc. ↑	0.84 ± 0.01	0.84 ± 0.01	0.79 ± 0.02	$\mathbf{0.95} \pm 0.00$	0.55 ± 0.01	$\underline{0.89} \pm 0.00$
	Div. →	6.72 ± 0.07	6.43 ± 0.03	$\underline{6.52} \pm 0.09$	6.47 ± 0.03	$\mathbf{6.84} \pm 0.10$	6.44 ± 0.01
	MMod. →	2.12 ± 0.07	$\mathbf{2.15} \pm 0.13$	2.35 ± 0.12	1.35 ± 0.07	4.03 ± 0.06	$\underline{2.04} \pm 0.03$
CMU	FID ↓	0.35 ± 0.04	$\underline{2.15} \pm 0.12$	2.46 ± 0.07	5.26 ± 0.13	4.61 ± 0.18	$\mathbf{1.11} \pm 0.06$
	Acc. ↑	0.85 ± 0.01	0.59 ± 0.02	$\mathbf{0.78} \pm 0.00$	0.28 ± 0.01	0.33 ± 0.01	$\underline{0.74} \pm 0.01$
	Div. →	6.28 ± 0.05	5.58 ± 0.11	5.39 ± 0.02	5.89 ± 0.04	$\mathbf{6.08} \pm 0.09$	$\underline{6.06} \pm 0.08$
	MMod. →	3.27 ± 0.08	4.68 ± 0.19	$\mathbf{2.84} \pm 0.21$	5.77 ± 0.05	5.65 ± 0.11	$\underline{4.24} \pm 0.14$

that require the SMPL body model. Evaluation uses a shared frozen action recognizer and lengths sampled by action from real data, capped at the smaller of the action's 95th-percentile real length and each generator's limit (Appendix A.2.2). Accuracy measures action agreement; diversity and multimodality measure between-sequence and within-action feature variation. Among these baselines, MotionSCM has the lowest Fréchet inception distance (FID) and second-highest accuracy on all three datasets. Its multimodality is among the two closest to Real (Table 3).

4.4 Structural Choices and Intervention Behaviour

Table 4 tests what the structural restrictions contribute beyond generation fit. A nearly parameter-matched direct decoder (691k versus 678k decoder parameters) obtains lower FID, but exposes no motion channel operators for mechanism intervention. Distributional fit alone therefore does not select an intervention-ready representation.

Gain control shapes both the learned coupling allocation and post-training intervention behaviour. Without gain constraints, the lagged cross-joint coefficient share rises from 5.5% to 70.6%, accompanied by larger attenuation responses and lower FID. Under a fivefold amplification request, the unconstrained model executes the edit directly and produces numerical explosions in 47.2% of 360-frame rollouts. MotionSCM projects the same request onto the admissible operator set and has no explosions in this test. The variant without DAG orientation retains gain constraints but removes the ordering mask, changing the final support density. It also achieves lower FID and smaller attenuation responses.

Coordination horizon. Longer windows lower FID, but the ratio of across-action to training-seed variation drops from 1.66 at $k = 5$ to 0.91 and 0.88 at $k = 9, 14$. Total coefficient mass remains nearly unchanged, but lagged mass shifts towards longer delays. The lower FID therefore

accompanies a different temporal allocation of coordination, not a more reproducible action-specific response. We use $k = 5$, where across-action variation exceeds training-seed variation.

Table 4: AIMove structural ablations over three training seeds (mean $\pm$ population SD). Response uses $0.5\times$ attenuation. Explosions are nonfinite states or normalized magnitude above 10 during 360-frame rollouts under requested $5\times$ amplification. Gain-constrained variants project requests; the unconstrained variant executes them directly. Dashes denote an unavailable structural intervention.

Variant	FID ↓	Cross-joint response	Explosion rate ↓
MotionSCM	0.476 ± 0.079	0.048 ± 0.010	0
Without gain constraints	0.413 ± 0.012	0.080 ± 0.012	0.472 ± 0.132
Without DAG orientation	0.363 ± 0.011	0.008 ± 0.002	0
Direct decoder	0.347 ± 0.009	—	—

5 Discussion

Within the learned generator, actions have characteristic coordination patterns with variation between motions that realize the same action. A rule's contribution depends on how its effects are transmitted and sustained. The fixed-target ablations show that lagged recurrence in receiving channels sustains perturbations transmitted across joints. In the tested single-seed budget settings, a larger lagged cross-joint share can accompany a smaller response (Appendix A.3.2). Coordination therefore needs to be analysed through the resulting dynamics, which coefficient shares alone do not characterize.

For generative modelling, these findings motivate evaluating mechanism control separately from distributional fit. Relaxed or direct decoders achieve lower FID without retaining the same intervention interface. Gain constraints govern amplification, while contemporaneous ordering gives the structural system a recursive form. The gain-constrained variant without DAG orientation has no numerical explosions in the stress test, illustrating that stability and recursive structure are distinct properties. Mechanism selectivity specifies which structural assignments are edited. Under matched target effects, smaller off-target responses indicate more localized editing.

Making these mechanisms intervenable provides a way to formulate hypotheses about human factors that alter movement organization. The fatigue-inspired edit examines the motion generated under a specified reduction in the transmission of coordination fluctuations. Linking this mechanism change to measured fatigue or cognitive load, and to task performance, requires human factor measurements and a separate causal inference study.

The causal interpretation applies to the learned generator under the stated assumptions. The fixed within-sequence contemporaneous ordering excludes within-frame feedback cycles, and gain constraints restrict admissible interventions. The same structural decoder could support text-conditioned or multimodal generation. Its known contemporaneous and lagged operators could also support semi-synthetic benchmarks for causal discovery and inference.

6 Conclusion

MotionSCM makes cross-joint coordination an explicit, intervenable part of human motion generation. Under stated causal assumptions, selective mechanism edits make an action's dependence on coordination rules directly testable. The structural model determines how a specified change affects the motion while generator inputs and unedited assignments remain fixed. This connects the study of motion coordination to the mechanisms through which an action is generated. MotionSCM retains competitive generation quality across three action-conditioned benchmarks, with rollout stability guaranteed for edits within the admissible operator set. It thereby extends motion generation to the controlled causal analysis of action organization.

AI USE STATEMENT

We used ChatGPT (OpenAI) and Claude (Anthropic) to provide feedback on experimental design, assist with implementation and debugging, check mathematical derivations and proof arguments, interpret results, and support manuscript drafting and revision. The authors tested the code, independently checked the mathematical arguments, and reviewed the interpretations and text. All research decisions were made by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Appendix A.1 provides the model specification, causal assumptions, and proofs. Appendix A.2 documents the datasets, preprocessing, training configuration, and evaluation protocols. Code and an inference checkpoint are available at `https://anonymous.4open.science/r/MotionSCM-736C/`.

REFERENCES

Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Jinseok Bae, Younghwan Lee, Donggeun Lim, and Young Min Kim. Plt: Part-wise latent tokens as adaptable motion priors for physically simulated characters. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Yiyang Cao, Yunze Deng, Ziyu Lin, Bin Feng, Xinggang Wang, Wenyu Liu, DanDan Zheng, and Jingdong Chen. Tric-motion: Tri-domain causal modeling grounded text-to-motion generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Patrick Chao, Patrick Blöbaum, Sapan Kirit Patel, and Shiva Kasiviswanathan. Modeling causal mechanisms with diffusion models for interventional and counterfactual queries. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Ling-Hao Chen, Wenxun Dai, Xuan Ju, Shunlin Lu, and Lei Zhang. Motionclr: Motion generation and training-free editing via understanding attention mechanisms. *arxiv:2410.18977*, 2024.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010. IEEE, 2023.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *International Conference on Learning Representations*, volume 2024, pp. 2923–2945, 2024.

Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 10093–10100, 2020.

Julie N Côté, Pierre A Mathieu, Mindy F Levin, and Anatol G Feldman. Movement reorganization to compensate for fatigue during sawing. *Experimental brain research*, 146(3):394–398, 2002.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74:981–995, 2007.

Dongjie Fu, Tengjiao Sun, Pengcheng Fang, Xiaohao Cai, and Hansung Kim. Mogo: Residual quantized hierarchical causal transformer for real-time and infinite-length 3d human motion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3994–4002, 2026.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *The Eleventh International Conference on Learning Representations*, 2023.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2021–2029, 2020.

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

Ziyan Guo, Zeyu Hu, De Wen Soh, and Na Zhao. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13869–13879, 2025.

Joshua Hanson and Maxim Raginsky. Universal simulation of stable dynamical systems by recurrent neural nets. In *Learning for Dynamics and Control*, pp. 384–392. PMLR, 2020.

Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7158–7168. IEEE, 2025.

Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. Causal discovery from subsampled time series data by constraint optimization. In *Conference on Probabilistic Graphical Models*, pp. 216–227. PMLR, 2016.

Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: from theory to practice. *Advances in Neural Information Processing Systems*, 36:58833–58864, 2023.

Biao Jiang, Xin Chen, Ailing Zeng, Xinru Sun, Fukun Yin, Xianfang Zeng, Xuanyang Zhang, Gang Yu, and Tao Chen. Causal motion tokenizer for streaming motion generation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2045–2055. IEEE, 2025a.

Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Dynamic motion blending for versatile motion editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22735–22745. IEEE, 2025b.

Aliasghar Khani, Arianna Rampini, Bruno Roy, Larasika Nadela, Noa Kaplan, Evan Atherton, Derek Cheung, and Jacky Bibliowicz. Motion generation: A survey of generative approaches and benchmarks. *arXiv preprint arXiv:2507.05419*, 2025.

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

Yonghyeon Lee, Byeongho Lee, Seungyeon Kim, and Frank C Park. Motion manifold flow primitives for task-conditioned trajectory generation under complex task-motion dependencies. *IEEE Robotics and Automation Letters*, 10(7):7412–7419, 2025.

Yuan-Ming Li, Qize Yang, Nan Lei, Shenghao Fu, Ling-An Zeng, Jian-Fang Hu, Xihan Wei, and Wei-Shi Zheng. Irg-motionllm: Interleaving motion generation, assessment and refinement for text-to-motion generation. *arXiv preprint arXiv:2512.10730*, 2025a.

Zekun Li, Sizhe An, Chengcheng Tang, Chuan Guo, Ivan Shugurov, Linguang Zhang, Amy Zhao, Srinath Sridhar, Lingling Tao, and Abhay Mittal. Llamo: Scaling pretrained language models for unified motion understanding and generation with continuous autoregressive tokens. *arXiv preprint arXiv:2602.12370*, 2026.

Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. Simmotionedit: Text-based human motion editing with motion similarity prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27827–27837. IEEE, 2025b.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023.

Helmut Lütkepohl et al. *New introduction to multiple time series analysis*, volume 1. springer Berlin, 2005.

Sotiris Manitsaris, Gavriela Senteri, Dimitrios Makrygiannis, and Alina Glushkova. Human movement representation on multivariate time series for recognition of professional gestures and forecasting their trajectories. *Frontiers in Robotics and AI*, 7:80, 2020.

Jose R Medina and Aude Billard. Learning stable task sequences from demonstration with linear parameter varying systems and hidden markov models. In *Conference on Robot Learning*, pp. 175–184. PMLR, 2017.

Brenda Elizabeth Olivas-Padilla, Alina Glushkova, and Sotiris Manitsaris. Motion capture benchmark of real industrial tasks and traditional crafts for human movement analysis. *IEEE Access*, 11:40075–40092, 2023.

Brenda Elizabeth Olivas-Padilla, Sotiris Manitsaris, and Alina Glushkova. Explainable ai in human motion: A comprehensive approach to analysis, modeling, and generation. *Pattern Recognition*, 151:110418, 2024.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pp. 1595–1605. Pmlr, 2020.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10965–10975. IEEE, 2021.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1546–1555. IEEE, 2024.

Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9955–9965, 2025.

Michele Rocca, Sune Darkner, Kenny Erleben, and Sheldon Andrews. Policy-space diffusion for physics-based character animation. *ACM Transactions on Graphics*, 44:1–18, 2025.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations*, volume 2024, pp. 8717–8733, 2024.

Keming Shen, Bizhu Wu, Junliang Chen, Xiaoqin Wang, and Linlin Shen. Finextrol: Controllable motion generation via fine-grained text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8905–8913, 2026.

Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pp. 1–48, 1980.

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

Kengo Uchida, Takashi Shibuya, Yuhta Takida, Naoki Murata, Julian Tanke, Shusuke Takahashi, and Yuki Mitsufuji. Mola: Motion generation and editing with latent diffusion enhanced by adversarial training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2901–2910. IEEE, 2025.

Mathukumalli Vidyasagar. *Nonlinear systems analysis*. SIAM, 2002.

Xinghan Wang, Kun Xu, Fei Li, Cao Sheng, Jiazhong Yu, and Yadong Mu. Generating attribute-aware human motions from textual prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10216–10224, 2026a.

Ziyi Wang, Xinshun Wang, Shuang Chen, Yang Cong, and Mengyuan Liu. Unimotion: A unified framework for motion-text-vision understanding and generation. *arXiv preprint arXiv:2603.22282*, 2026b.

Dongze Wu, Feng Qiu, and Yao Xie. Doflow: Flow-based generative models for interventional and counterfactual forecasting on time series. In *International Conference on Learning Representations*, volume 2026, pp. 149949–149980, 2026.

Yutong Xia, Chang Xu, Yuxuan Liang, Li Zhao, Qingsong Wen, Roger Zimmermann, and Jiang Bian. Causal time series generation via diffusion models. *arXiv preprint arXiv:2509.20846*, 2025.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *International Conference on Learning Representations*, volume 2024, pp. 28176–28194, 2024.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9588–9597. IEEE, 2021.

Qing Yu, Akihisa Watanabe, and Kent Fujiwara. Causal motion diffusion models for autoregressive motion generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14730–14740, 2023a.

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36:13981–13992, 2023b.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.

Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pp. 18–38. Springer, 2024.

Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. In *European conference on computer vision*, pp. 126–143. Springer, 2024.

A APPENDIX

CONTENTS

A.1 Method Details 14
A.1.1 SCM Construction Conditions and Causal Assumptions 14
A.1.2 Operator Parameterization and Gain Constraints 15
A.1.3 Acyclicity Projection 15
A.1.4 Budgets, Reserve, and Projection 16
A.1.5 Proofs 17
A.1.6 From the Represented State to Physical Units 18
A.2 Experimental Details 19
A.2.1 Datasets and Preprocessing 19
A.2.2 Evaluation Protocol 19
A.2.3 Implementation Details 22
A.3 Additional Results and Diagnostics 22
A.3.1 Mechanism Intervention Diagnostics 22
A.3.2 Structural and Temporal Ablations 24
A.3.3 Editor Comparison Details 27
A.3.4 Generation Diagnostics and Qualitative Results 28

A.1 METHOD DETAILS

A.1.1 SCM CONSTRUCTION CONDITIONS AND CAUSAL ASSUMPTIONS

The structural decoder induces a generator-level SCM over the rollout states, $M_\theta = (U, V, F_\theta, P_U)$. The endogenous variables are the generated states $V = \{x_{t,j} : t > k\}$, while $U = (c, \mathbf{z}_{1:T'}, \mathbf{x}_{1:k}, \mathbf{b}_{1:T})$ contains the conditioning signal, sampled latent trajectory, seed frames, and motion drive as exogenous generator variables. We write $\mathcal{G}_\theta$ for the decoder graph and $\mathrm{pa}_\theta(t, j)$ for the parents of $x_{t,j}$.

Row i of Eq. (1) is $x_{t,i} = \sum_j \mathbf{A}_0^{(t)}[i,j]x_{t,j} + \sum_{\ell=1}^{k}\sum_j \mathbf{A}_\ell^{(t)}[i,j]x_{t-\ell,j} + b_{t,i}$. This is the channel-level structural assignment used by the intervention interface.

Construction conditions.

- **C1 Structural assignments.** Each endogenous variable obeys $x_{t,j} = f_{\theta,t,j}(\mathbf{x}_{\mathrm{pa}_\theta(t,j)}, U)$, corresponding to Eq. (1) written one channel at a time.
- **C2 Per-sequence acyclic ordering.** The coefficient decoders depend on $(\mathbf{z}, c, r_t)$ and not on the generated state, so each realization $U = u$ fixes the contemporaneous operators and a total order $\prec_u$ over the D channels independently of the rollout states. Edges satisfy $x_{s,i} \to x_{t,j} \Rightarrow s < t$ or $(s = t \wedge i \prec_u j)$, making $\mathbf{A}_0^{(t)}$ strictly triangular and $\mathbf{I} - \mathbf{A}_0^{(t)}$ invertible at every frame.

Causal assumptions.

- **A1 Generator-level causal sufficiency.** Within the generator, all common causes of the endogenous variables are included in U.
- **A2 Exogeneity.** The variables in U have no endogenous parents, and their distribution P_U is unchanged by interventions on V.

A3 Modularity. A mechanism intervention $\mathrm{do}(\mathbf{A}_S := \widetilde{\mathbf{A}}_S)$ replaces only the selected structural assignments. Writing $F_\theta = (f_S, f_{\bar{S}})$, the intervention maps F_θ to $(f'_S, f_{\bar{S}})$ while leaving P_U and the unedited assignments fixed.

The decoder supports mechanism (soft) and value (atomic) interventions with standard SCM semantics (Pearl, 2009; Eberhardt & Scheines, 2007; Correa & Bareinboim, 2020).

Conditional on $U = u$, C1 and C2 define a recursive SCM. Under A1–A3, mechanism and value interventions have the standard SCM semantics (Pearl, 2009; Peters et al., 2017). An intervention or unit-level counterfactual is obtained by re-solving Eq. (1) at the same u, with the edited assignments substituted and all unedited assignments held fixed.

A.1.2 Operator Parameterization and Gain Constraints

For channel i, the lagged part of its structural assignment is written as a direction multiplied by a scalar gain. Let $\mathbf{C}_\ell^{(t)}$ denote the coefficient basis mixture at lag ℓ. We normalize the row jointly over all lags and source channels,

$$\mathbf{d}_{\ell,i}^{(t)} = \frac{\mathbf{C}_\ell^{(t)}[i,:]}{\sum_{\ell'=1}^{k} \|\mathbf{C}_{\ell'}^{(t)}[i,:]\|_1},$$

and assign it the learned gain

$$\rho_i^{(t)} = q\,\tau_{\mathrm{lag}}\,\sigma(g_{t,i}). \tag{2}$$

where σ is the logistic function, $\tau_{\mathrm{lag}} \in (0,1)$ is the lagged coupling budget, and $q \in (0,1)$ leaves reserve for intervention. The lagged operators are then $\mathbf{A}_\ell^{(t)}[i,:] = \rho_i^{(t)}\mathbf{d}_{\ell,i}^{(t)}$, giving

$$\sum_{\ell=1}^{k} \|\mathbf{A}_\ell^{(t)}[i,:]\|_1 \le \rho_i^{(t)} < q\,\tau_{\mathrm{lag}}.$$

The normalization uses $\max\{\cdot, 10^{-6}\}$ in the denominator, so the bound holds with equality whenever the normalizing sum clears that floor, and a row whose basis mixture vanishes stays zero instead of being renormalized. Thus the distribution of lagged coupling across source channels and delays is learned independently of its total gain.

The contemporaneous operator $\mathbf{A}_0^{(t)}$ uses a separate rowwise L_1 budget. After acyclicity is enforced, let $R_i = \sum_j |\mathbf{A}_0^{(t)}[i,j]|$. Each row is rescaled as

$$\mathbf{A}_0^{(t)}[i,:] \leftarrow \mathbf{A}_0^{(t)}[i,:]\,\min\{1, \kappa/R_i\},$$

so $\|\mathbf{A}_0^{(t)}\|_\infty \le \kappa$. Training and ordinary generation use $\kappa = q\,\tau_{\mathrm{con}}$, while the admissible post-intervention ceiling is τ_{con}, leaving explicit reserve for amplification.

Because the acyclic construction makes $\mathbf{I}_D - \mathbf{A}_0^{(t)}$ invertible, we define the solved operators

$$\mathbf{F}_\ell^{(t)} = (\mathbf{I}_D - \mathbf{A}_0^{(t)})^{-1}\mathbf{A}_\ell^{(t)}, \qquad \mathbf{u}_t = (\mathbf{I}_D - \mathbf{A}_0^{(t)})^{-1}\mathbf{b}_t,$$

which give the explicit recursion

$$\mathbf{x}_t = \sum_{\ell=1}^{k} \mathbf{F}_\ell^{(t)}\mathbf{x}_{t-\ell} + \mathbf{u}_t. \tag{3}$$

A.1.3 Acyclicity Projection

To give the contemporaneous effects a recursive SCM interpretation, $\mathbf{A}_0^{(t)}$ must be acyclic at every frame. We construct one support and one total order for each sequence from the time-averaged raw coefficients, and apply them to every frame. Let $\mathbf{C}_0^{(t)}[i,j]$ denote the raw coefficient of the edge from source j to target i.

Algorithm 1 Per-sequence acyclicity projection

Require: Raw coefficients $\{\mathbf{C}_0^{(t)} \in \mathbb{R}^{D\times D}\}_{t=1}^{T}$
Require: Keep fraction $\gamma = 0.10$ and magnitude floor $\eta = 10^{-4}$
Ensure: Acyclic operators $\{\mathbf{A}_0^{(t)}\}_{t=1}^{T}$

$\bar{\mathbf{C}} \leftarrow T^{-1}\sum_{t=1}^{T} |\mathbf{C}_0^{(t)}|$; set $\bar{\mathbf{C}}[i,i] \leftarrow 0$
$K \leftarrow \max\{1, \lceil \gamma D(D-1) \rceil\}$
$\xi \leftarrow \max\{\mathrm{KthLargest}(\mathrm{vec}(\bar{\mathbf{C}}), K), \eta\}$
$E \leftarrow \{(i,j) : i \neq j \ \wedge \ \bar{\mathbf{C}}[i,j] \geq \xi\}$
for $j = 1, \ldots, D$ **do**
    $w_j \leftarrow \sum_{i:(i,j)\in E} \bar{\mathbf{C}}[i,j] - \sum_{i:(j,i)\in E} \bar{\mathbf{C}}[j,i]$
end for
$\prec \leftarrow$ channels sorted by w_j in descending order
for $t = 1, \ldots, T$ **do**
    for all $i, j \in \{1, \ldots, D\}$ **do**
        if $(i,j) \in E \ \wedge \ j \prec i$ **then**
            $\mathbf{A}_0^{(t)}[i,j] \leftarrow \mathbf{C}_0^{(t)}[i,j]$
        else
            $\mathbf{A}_0^{(t)}[i,j] \leftarrow 0$
        end if
    end for
end for
return $\{\mathbf{A}_0^{(t)}\}_{t=1}^{T}$

The time-averaged coefficients determine a shared support, while the net outgoing-minus-incoming score w_j defines the channel order. Retaining only edges that follow this order makes $\mathbf{A}_0^{(t)}$ strictly triangular at every frame, so the contemporaneous graph is acyclic and $\mathbf{I}_D - \mathbf{A}_0^{(t)}$ is invertible. Training applies the hard projection in the forward pass and passes gradients through unchanged in the backward pass. The subsequent row-L_1 cap preserves its support and ordering.

A.1.4 Budgets, Reserve, and Projection

Budget and reserve. Proposition 1 requires the executed operators to satisfy the lagged and contemporaneous budgets τ_{lag} and τ_{con}. The decoder instead emits within $q\,\tau_{\mathrm{lag}}$ and $q\,\tau_{\mathrm{con}}$, with $q < 1$ leaving reserve for amplification. We use $\tau_{\mathrm{lag}} = 0.64$, $\tau_{\mathrm{con}} = 0.35$, and $q = 2/3$, so $\tau_{\mathrm{lag}} + \tau_{\mathrm{con}} = 0.99 < 1$. From Eq. (2), a lagged row admits

$$a_i^{\max,(t)} = \frac{\tau_{\mathrm{lag}}}{\rho_i^{(t)}} = \frac{1}{q\,\sigma(g_{t,i})} \geq \frac{1}{q},$$

and the same reserve argument applies to the contemporaneous row cap. Thus every emitted row with nonzero lagged mass can be amplified by at least $1/q = 1.5$ without exceeding the coupling budgets.

Projection. After an edit, the projection Π enforces the full budgets independently for each target row. Row i of the contemporaneous operator is rescaled by

$$\min\left\{1, \frac{\tau_{\mathrm{con}}}{\|\widetilde{\mathbf{A}}_0[i,:]\|_1}\right\},$$

and its k lagged slices are rescaled jointly by

$$\min\left\{1, \frac{\tau_{\mathrm{lag}}}{\sum_{\ell=1}^{k} \|\widetilde{\mathbf{A}}_\ell[i,:]\|_1}\right\}.$$

Zero rows are left unchanged. In particular, the first channel in the acyclic order has no contemporaneous parents. The projection enforces the row budgets and not the ordering $\prec_u$, so a contemporaneous edit keeps that ordering by construction: editing coefficients on the existing support is sufficient,

and is what every intervention reported here does. The projection therefore leaves edits inside the budgets unchanged, modifies only rows that exceed them, and preserves the support of the contemporaneous operator. The AIMove attenuation interventions and the MotionSCM interventions in the matched editor comparison execute the requested operators without further rescaling, with every unedited row unchanged. The amplification tests in Appendix A.3.2 also evaluate requests beyond the budgets.

The same uniform gain bound applies to simultaneous or time-varying mechanism interventions when every executed row satisfies these budgets and the contemporaneous order is preserved. For repeated edits, the budgets apply to the resulting rows, and the $1/q$ amplification reserve is measured relative to the original decoder output.

A.1.5 PROOFS

Proof of Proposition 1. Uniform gain. Fix a frame t and omit the superscript (t) for readability. Since $\|\mathbf{A}_0\|_\infty \le \tau_{\mathrm{con}} < 1$, the Neumann bound gives

$$\|(\mathbf{I}-\mathbf{A}_0)^{-1}\|_\infty \le \frac{1}{1-\|\mathbf{A}_0\|_\infty} \le \frac{1}{1-\tau_{\mathrm{con}}}.$$

Using $\mathbf{F}_\ell = (\mathbf{I}-\mathbf{A}_0)^{-1}\mathbf{A}_\ell$, we have entrywise

$$\sum_{\ell=1}^{k} |\mathbf{F}_\ell| \le |(\mathbf{I}-\mathbf{A}_0)^{-1}| \sum_{\ell=1}^{k} |\mathbf{A}_\ell|.$$

Taking the infinity norm and applying the lagged row gain bound gives

$$G_t \le \|(\mathbf{I}-\mathbf{A}_0)^{-1}\|_\infty \left\| \sum_{\ell=1}^{k} |\mathbf{A}_\ell| \right\|_\infty \le \frac{\tau_{\mathrm{lag}}}{1-\tau_{\mathrm{con}}} = g_\star.$$

Because $\tau_{\mathrm{lag}} + \tau_{\mathrm{con}} < 1$, we have $g_\star < 1$. Since the argument holds for every t, $\bar{g} \le g_\star < 1$.

Window decay. Fix a starting frame s and define the maximum deviation in the j-th k-frame block as

$$M_j = \max_{0 \le r < k} \|\boldsymbol{\delta}_{s+jk+r}\|_\infty.$$

For the first frame of block $j+1$, all k preceding states lie in block j, so the deviation recursion gives $\|\boldsymbol{\delta}_{s+(j+1)k}\|_\infty \le \bar{g} M_j$. Proceeding through the block, each new frame depends only on states from block j or on earlier frames of block $j+1$, which are already bounded by $\bar{g} M_j \le M_j$. Hence every frame in block $j+1$ is bounded by $\bar{g} M_j$, and therefore

$$M_{j+1} \le \bar{g} M_j.$$

Iterating gives $M_j \le \bar{g}^{\,j} M_0$. For $h = jk + r$ with $0 \le r < k$,

$$\|\boldsymbol{\delta}_{s+h}\|_\infty \le M_j \le \bar{g}^{\,j} M_0 = \bar{g}^{\,\lfloor h/k \rfloor} \max_{0 \le \ell < k} \|\boldsymbol{\delta}_{s+\ell}\|_\infty,$$

which proves the claimed decay bound.

□

Proof of Corollary 1. Fix the exogenous realization and define

$$X_0 = \max_{1 \le j \le k} \|\mathbf{x}_j\|_\infty, \qquad B = \sup_{t \le T} \|\mathbf{u}_t\|_\infty < \infty.$$

For the original rollout, let

$$X = \max\left\{ X_0, \frac{B}{1-\bar{g}} \right\}.$$

Proof. Bounded state. The seed frames satisfy $\|\mathbf{x}_j\|_\infty \le X$ by definition. Suppose all preceding states are bounded by X. From Eq. (3),

$$\|\mathbf{x}_t\|_\infty \le \bar{g} \max_{1 \le \ell \le k} \|\mathbf{x}_{t-\ell}\|_\infty + B \le \bar{g} X + B \le X,$$

where the last inequality follows from $X \geq B/(1-\bar{g})$. Induction therefore gives $\|\mathbf{x}_t\|_\infty \leq X$ for every $t \leq T$.

Mechanism intervention. Let $\widetilde{\mathbf{F}}_\ell^{(t)}$ and $\widetilde{\mathbf{u}}_t$ denote the solved operators and drive after the executed intervention, and define

$$\widetilde{g} = \sup_{t \leq T} \left\| \sum_{\ell=1}^{k} |\widetilde{\mathbf{F}}_\ell^{(t)}| \right\|_\infty \leq g_\star < 1.$$

The magnitude of the induced change in the solved dynamics is measured by

$$\Delta_F = \max_{t \leq T} \max_i \sum_{\ell=1}^{k} \left\| \widetilde{\mathbf{F}}_\ell^{(t)}[i,:] - \mathbf{F}_\ell^{(t)}[i,:] \right\|_1, \qquad \Delta_u = \max_{t \leq T} \|\widetilde{\mathbf{u}}_t - \mathbf{u}_t\|_\infty.$$

Because the two rollouts share the same seed frames, $\boldsymbol{\delta}_t = \widetilde{\mathbf{x}}_t - \mathbf{x}_t = \mathbf{0}$ on the seed. Subtracting their recursions gives

$$\boldsymbol{\delta}_t = \sum_{\ell=1}^{k} \widetilde{\mathbf{F}}_\ell^{(t)} \boldsymbol{\delta}_{t-\ell} + \sum_{\ell=1}^{k} \left(\widetilde{\mathbf{F}}_\ell^{(t)} - \mathbf{F}_\ell^{(t)} \right) \mathbf{x}_{t-\ell} + (\widetilde{\mathbf{u}}_t - \mathbf{u}_t).$$

Using the bound $\|\mathbf{x}_t\|_\infty \leq X$,

$$\|\boldsymbol{\delta}_t\|_\infty \leq \widetilde{g} \max_{1 \leq \ell \leq k} \|\boldsymbol{\delta}_{t-\ell}\|_\infty + \Delta_F X + \Delta_u.$$

Starting from the shared seed and applying the same induction as above yields

$$\max_{t \leq T} \|\boldsymbol{\delta}_t\|_\infty \leq \frac{\Delta_F X + \Delta_u}{1 - \widetilde{g}}.$$

Although the raw motion drive $\mathbf{b}_t$ is held fixed under a mechanism intervention, Δ_u can be nonzero when $\mathbf{A}_0^{(t)}$ is edited because $\mathbf{u}_t = (\mathbf{I} - \mathbf{A}_0^{(t)})^{-1}\mathbf{b}_t$. For an intervention confined to the lagged operators, $\Delta_u = 0$. □

A.1.6 From the Represented State to Physical Units

Proposition 1 and Corollary 1 apply to the normalized state $\mathbf{x}_t$. For HumanAct12 and CMU MoCap, each channel represents one Cartesian joint coordinate; for AIMove, the channels represent joint angles except for the Hips yaw increment. All channels are mapped from normalized coordinates by the fixed affine transformation

$$y_{t,j} = s_j x_{t,j} + m_j, \qquad s_j = \frac{x_{\mathrm{hi},j} - x_{\mathrm{lo},j}}{2}, \qquad m_j = \frac{x_{\mathrm{hi},j} + x_{\mathrm{lo},j}}{2}, \tag{4}$$

where $x_{\mathrm{lo},j}$ and $x_{\mathrm{hi},j}$ are the per-channel clipping bounds fitted on the training split.

Directly represented channels. The absolute joint angle channels of AIMove and all Cartesian coordinate channels of HumanAct12 and CMU MoCap are represented directly in physical coordinates after the affine map.

$$|y_{t,j} - \widetilde{y}_{t,j}| = s_j |\delta_{t,j}| \leq s_{\max} \|\boldsymbol{\delta}_t\|_\infty, \qquad s_{\max} = \max_{j \text{ direct}} s_j.$$

Therefore Corollary 1 gives

$$|y_{t,j} - \widetilde{y}_{t,j}| \leq s_{\max} \frac{\Delta_F X + \Delta_u}{1 - \widetilde{g}}$$

for a mechanism intervention, while the decay bound of Proposition 1 transfers analogously to a value intervention.

Integrated AIMove heading. On AIMove, the Hips yaw channel represents a per-frame increment, and the physical heading is recovered by cumulative sum,

$$\theta_T = \theta_0 + \sum_{t=1}^{T} (s_\theta x_{t,\theta} + m_\theta).$$

For a mechanism intervention, the offset and initial heading cancel between the two rollouts, giving

$$|\theta_T - \widetilde{\theta}_T| \le s_\theta \sum_{t=1}^{T} \|\boldsymbol{\delta}_t\|_\infty \le s_\theta T \frac{\Delta_F X + \Delta_u}{1 - \widetilde{g}}.$$

Thus the increment remains uniformly bounded, while the heading has a finite-horizon bound that grows linearly with the horizon because it is obtained by cumulative sum. For a value intervention, Proposition 1 gives a decaying deviation, so

$$|\theta_{s+h} - \widetilde{\theta}_{s+h}| \le \frac{s_\theta k}{1 - \bar{g}} \max_{0 \le \ell < k} \|\boldsymbol{\delta}_{s+\ell}\|_\infty,$$

which is uniform in h.

A.2 Experimental Details

Code availability. Code and inference checkpoints: `https://anonymous.4open.science/r/MotionSCM-736C/`.

A.2.1 Datasets and Preprocessing

AIMove. Each 90 Hz BVH sequence is represented by 81 unwrapped rotation channels from 27 joints. We discard root translation, express the Hips rotations relative to the first frame, and represent Hips yaw as per-frame increments (Appendix A.1.6). Starting from 1079 raw captures, we exclude 19 captures flagged in the quality manifest for duration or label errors, one malformed skeleton, and six captures with incorrect frame rates. We also remove 46 within-label duration outliers using the threshold $\min\{Q_3 + 3\,\mathrm{IQR}, \mathrm{median} + 4\,\mathrm{MAD}\}$, followed by five sequences from labels left with fewer than three examples. The resulting corpus contains 1002 sequences.

The source annotation hierarchy defines nine candidate groups for pooling fine-grained labels. We merge a group only when the largest Fisher ratio computed over duration, channel variability, leg ROM, and root yaw excursion is below 1.5. Eight of the nine groups satisfy this criterion. In the remaining group, GBBC, GBFD, and GBTP are retained as separate labels. This produces the fixed 25-class mapping used throughout the paper, with exact membership recorded in the preprocessing manifest.

Public benchmarks. For HumanAct12 and CMU MoCap, we use the processed data distributed with Guo et al. (2020) without additional skeleton preprocessing. Two CMU sequences shorter than the k-frame seed are excluded from model fitting but retained in the full real reference used for evaluation.

Normalization and splits. We clip each channel at quantiles estimated from the training split, then map it affinely to $[-1, 1]$. We use Q_1/Q_{99} for AIMove pose channels and $Q_{0.1}/Q_{99.9}$ for its yaw increment and all position channels. Using seed 42, we split sequences within each class and hold out 15% for validation on all datasets. Only the training split is used to fit the models and normalization statistics, while the validation split is reserved for held-out diagnostics. Distributional generation metrics use the full real corpus as their reference. Table A1 summarizes the resulting representations, sample counts, and evaluation windows.

A.2.2 Evaluation Protocol

Generation metrics. All methods are evaluated with a fixed action recognizer for each dataset. We use the released GRU classifiers of Guo et al. (2020) for HumanAct12 and CMU MoCap and train the same architecture on the AIMove training split. Feature extraction follows Petrovich et al.

Table A1: Dataset summary and generation evaluation protocol. Train/val. denotes the fixed model fitting split, while Real denotes the full corpus used as the distributional reference.

	AIMove	HumanAct12	CMU MoCap
Input	27 joints / 81 rotations	24 joints / 72 positions	20 joints / 60 positions
Frame rate	90 Hz	released data	12 Hz
Train/val.	851/151	1012/179	924/162
Real / n_{FID}	1002/501	1191/595	1088/544
Evaluation window	200	60	100

(2021). Recognizer inputs are root-centered at the first frame, with CMU coordinates additionally divided by 20 following the released evaluator. Sequences longer than the evaluation window in Table A1 are uniformly cropped, while shorter sequences retain their original length; the same rule is applied to real and generated motion.

Generated actions follow the class proportions of the real corpus, and each method produces at least n_{FID} sequences per seed. Target lengths are sampled from the empirical class-conditional length distribution of the real corpus and capped at the smaller of that class's 95th-percentile real length and the maximum length supported by each generator. The same length-sampling protocol is used for all methods. For fixed-length generators, the generated sequence is truncated to the sampled target length before evaluation.

We use a common FID sample size of $n_{\mathrm{FID}} = \lfloor N_{\mathrm{real}}/2 \rfloor$. For each seed, the reported FID is the median of 200 estimates, each computed between a random half of the real corpus and n_{FID} generated sequences sampled with replacement. Accuracy measures agreement between the conditioning action and the recognizer. Diversity is the average of 200 random between-sequence distances in recognizer feature space. Multimodality is computed from 20 within-action pairs and then averaged across actions, using the full generated pool. Real entries are obtained from 20 pairs of size-n_{FID} bootstrap samples and report mean $\pm 1.96 s/\sqrt{20}$. Method entries report mean $\pm$ population standard deviation over three independent training seeds. The AIMove MotionSCM runs use the Stage 1 checkpoint with the lowest validation reconstruction loss and the final Stage 2 checkpoint.

Baselines. ACTOR, MDM, MLD, and EMDM are retrained on the same training splits using their released implementations. Adaptations align the input/output layers and reconstruction objectives with each dataset's motion channels while retaining the core generative architecture and sampler. SMPL-dependent losses and conversions are omitted for these channel representations. Sampling uses one decoder pass for ACTOR, 1000 DDPM steps for MDM, 50 DDIM steps for MLD, and 10 DDGAN steps for EMDM. All models use the common evaluation protocol above. Unless otherwise specified, baseline training and generation settings follow the defaults in their respective official repositories.

Intervention metrics. Let $\mathbf{x}^{\mathrm{b}}, \mathbf{x}^{\mathrm{e}} \in \mathbb{R}^{T \times D}$ denote base and edited rollouts generated with the same conditioning action, latent trajectory, seed frames $\mathbf{x}_{1:k}$, and raw drive $\mathbf{b}_t$. For channel c, we define

$$\mathrm{ROM}_c(\mathbf{x}, W) = \max_{t \in W} x_{t,c} - \min_{t \in W} x_{t,c}.$$

ROM is evaluated in normalized motion channels with a denominator floor of $\varepsilon = 10^{-6}$, inactive for all 192,000 unedited channel–sample entries in Figure 5. The response of channel c is

$$r_c = \frac{\left|\mathrm{ROM}_c(\mathbf{x}^{\mathrm{e}}, W_{\mathrm{r}}) - \mathrm{ROM}_c(\mathbf{x}^{\mathrm{b}}, W_{\mathrm{r}})\right|}{\max\{\mathrm{ROM}_c(\mathbf{x}^{\mathrm{b}}, W_{\mathrm{r}}), \varepsilon\}}, \qquad W_{\mathrm{r}} = \{k+1, \ldots, 5k\},$$

measured over the first $4 \times k$ solved frames. For an intervention on channel i, the cross-joint response is the mean of r_c over all $c \neq i$. The edited-channel response is

$$\frac{\mathrm{ROM}_i(\mathbf{x}^{\mathrm{e}}, W_{\mathrm{s}})}{\max\{\mathrm{ROM}_i(\mathbf{x}^{\mathrm{b}}, W_{\mathrm{s}}), \varepsilon\}}, \qquad W_{\mathrm{s}} = \{5k+1, \ldots, T\}.$$

The whole-row attenuation and per-action response analyses in Section 4.2 use $T = 180$, 32 prior draws per action, all 25 actions, and three training seeds. We average responses over draws and actions within each training seed, then report their mean and population standard deviation. For per-action analyses, let $\rho_{s,a}$ denote the draw-averaged cross-joint response for training seed s and action a. In Table A4, σ_{action} is the across-action standard deviation of the seed-averaged responses, and σ_{seed} is the across-seed standard deviation averaged over actions. Action agreement is measured with the AIMove recognizer over the full 180-frame rollout. Relative ROM is unchanged by a fixed per-channel affine map with nonzero scale when the denominator floor is inactive in both coordinate systems. ROM measures amplitude change; trajectory RMS complements it in the propagation and editor comparisons. The editor comparison uses common Cartesian coordinates.

Fatigue-inspired intervention. The intervention in Section 4.2 attenuates coupling fluctuations around the forearm row's base-rollout mean while retaining the raw drive:

$$\mathrm{do}(f_{\theta,t,i} := \bar{c}_i + \alpha(c_{t,i} - \bar{c}_i) + b_{t,i}).$$

Here, $c_{t,i}$ is the coupling input, $\bar{c}_i$ its mean after the first 40 solved frames of the base rollout, and α the retained fluctuation strength. We use $\alpha \in \{0.8, 0.5, 0.2\}$, four prior draws per action, all 25 actions, and three training seeds, with $T = 180$. Figure 6 reports one minus the edited-channel ROM ratio over W_{s}, whose denominator is floored at 10^{-9}; unedited response uses W_{r} and $\varepsilon = 10^{-6}$ as above.

Editor comparison. Table 2 compares walking, running, and boxing using five independently generated sources per method and action (sampling seeds 0–4). Figure 1 shows the walking subset with seeds $0, 1, 2$. MotionCLR and SALAD use their released HumanML3D checkpoints with the prompts "a person walks forward swinging both arms", "a person runs forward swinging both arms", and "a person is boxing and punching with both arms"; MotionSCM uses its HumanAct12 checkpoint with the corresponding action labels. For each source, the latent input or sampling noise is held fixed throughout the control search. The comparison matches the edited body region and the relative magnitude of its response across the native sources of the three methods.

Table A2: Native controls and calibrated strengths for sampling seeds $0, 1, 2$, respectively.

Method	Native control	Calibrated strengths
MotionCLR	Attention offset for "both arms"	$-0.01,\ -0.00375,\ -0.015$
SALAD	Attention multiplier for "arms"	$0.625,\ 0.84375,\ 0.5625$
MotionSCM	Bilateral-arm coupling-row multiplier	$0.984375,\ 0.984375,\ 0.9875$

We evaluate 80-frame sequences on a common set of 22 corresponding SMPL landmarks, using frames 6–80 and expressing each rollout relative to its own pelvis. The target set S contains the bilateral collar, shoulder, elbow, and wrist joints, while the remaining 13 non-pelvis joints form the untargeted set U. For MotionSCM, we use the 22 landmarks corresponding to the HumanML3D/SMPL joint set. Its intervention scales the incoming Cartesian coordinate rows of all eight target joints across the contemporaneous and lagged operators while leaving the remaining structural rows unchanged.

Let r_j denote the RMS displacement between the base and edited pelvis-relative trajectories of joint j over the evaluation frames,

$$r_j = \left(\frac{1}{T_{\mathrm{e}}} \sum_t \left\| \mathbf{p}^{\mathrm{e}}_{t,j} - \mathbf{p}^{\mathrm{b}}_{t,j} \right\|_2^2 \right)^{1/2},$$

where T_{e} is the number of evaluated frames. For a joint set G, we define

$$r_G = \left(\frac{1}{|G|} \sum_{j \in G} r_j^2 \right)^{1/2}.$$

Let a_S denote the RMS excursion of the target joints in the unedited source motion about their temporal means. For each source, calibration targets $r_S/a_S \in [0.09, 0.11]$ using the same expansion-and-bisection search, with at most 12 non-neutral control evaluations. Table A2 reports the control strengths for the three-source walking subset.

Figure 1 maps the mean jointwise responses r_j/r_S to a shared reference pose. The accompanying bars report the mean normalized response r_G/r_S for the five trunk/head joints and the eight leg joints, using a common scale from zero to 1.2.

A.2.3 Implementation Details

Architecture. Stage 1 uses a four-layer Transformer encoder of width 256 with eight heads to produce $T' = 32$ latent tokens of dimension $d_z = 32$. The SeedDecoder and DriveDecoder are one-hidden-layer MLPs of width 128, with drive rank $R = 32$. The shared CoefDecoder has two width-128 hidden layers; its CoefHead branch produces $M = 16$ unit-normalized mixture weights and its GainHead branch produces D gain logits. We use $k = 5$ and the budgets in Appendix A.1.4. The Stage 2 prior is a six-layer Transformer of width 192 with four heads, learned token positions, flow time embeddings, and conditional/null-class embeddings for guidance.

Flow matching prior. Stage 2 fits a conditional flow matching prior $p_\psi(\mathbf{z}_{1:T'} \mid c)$ to the inferred latent trajectories. Its velocity field v_ψ is trained with $\mathbb{E}\,\|v_\psi(\mathbf{z}_{t_\mathrm{f}}, t_\mathrm{f}, c) - (\mathbf{z} - \boldsymbol{\epsilon})\|_2^2$. Here, $t_\mathrm{f} \in [0, 1]$ is the flow matching time, $\mathbf{z}_{t_\mathrm{f}} = (1 - t_\mathrm{f})\boldsymbol{\epsilon} + t_\mathrm{f}\mathbf{z}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{z}$ is a posterior latent trajectory from Stage 1.

Training. Stage 1 minimizes $\mathcal{L}_\mathrm{rec} + \beta\mathcal{L}_\mathrm{KL} + 0.1\mathcal{L}_\mathrm{res} + \mathcal{L}_\mathrm{seed} + 0.1\mathcal{L}_\mathrm{stab}$. The terms are, respectively, free-rollout MSE initialized by the preceding real k-frame seed, posterior KL, one-step residual Gaussian NLL, seed reconstruction MSE, and $\mathbb{E}[\max\{0, g_\mathrm{lag} - 0.9\}]$ for maximum lagged row gain g_lag. The rollout spans a random chunk of at most 128 frames and perturbs its seed with Gaussian noise of standard deviation 0.01. The KL weight rises from zero to 10^{-3} over the first ten epochs of each 20-epoch cycle. Stage 2 targets are fresh posterior samples, standardized with training statistics and smoothed over three latent tokens; null-class dropout is 0.1. Both stages use AdamW and gradient clipping at 1.0. Stage 1 uses learning rate 10^{-4}, weight decay 10^{-5}, and batch size 1; Stage 2 uses 10^{-3}, 10^{-4}, and 64. AIMove trains for 300/400 Stage 1/Stage 2 epochs, while HumanAct12 and CMU MoCap use 80/4000.

Generation. We integrate the latent flow for 100 Euler steps and decode the requested length. Classifier-free guidance uses $w = 2.0$ on HumanAct12 and $w = 2.5$ on CMU MoCap; AIMove uses the conditional field ($w = 1$). Intervention latents come from the same prior. Each model is trained on one NVIDIA RTX 3500 Ada laptop GPU with 12 GB memory.

Structural decoding cost. Before rollout, the implementation computes the inverses of $\mathbf{I}_D - \mathbf{A}_0^{(t)}$ in a batch across frames and multiplies them into the lagged operators and drive. For dense operators, this stage costs $O(T(k+1)D^3)$, followed by $O(TkD^2)$ for the recursion. Storing the operators requires $O(T(k+1)D^2)$ memory. These operator processing costs are additional to the neural decoder and latent prior.

A.3 Additional Results and Diagnostics

A.3.1 Mechanism Intervention Diagnostics

Whole-row intervention and action agreement. Applying the 0.5 attenuation of Section 4.2 only to the contemporaneous row reduces cross-joint response from 0.048 ± 0.010 to 0.021 ± 0.004 over the same three training seeds and 25 actions. Real validation sequences cropped to the same 180-frame length have action agreement 0.78, providing a reference for the base and edited values reported in Section 4.2. This readout measures preservation of action labels. For most actions, the legs show the largest unedited-channel responses and the head responds weakly.

Table A3 checks sensitivity to small base ROM in the same 25-action analysis. For each model, we pool base ROM over actions, draws, and unedited channels on W_r, then exclude channel–sample

entries below the 10th or 25th percentile. We average retained channels within each draw before averaging draws and actions. Action rankings remain correlated with the unfiltered readout, while response magnitude varies with this filtering.

Table A3: Sensitivity of the Figure 5 readout to base-ROM filtering. Response is mean ± population standard deviation over three training seeds. Spearman ranges compare the 25 action means with the unfiltered readout within each seed.

Excluded base-ROM entries	Cross-joint response	$\sigma_{\text{action}}/\sigma_{\text{seed}}$	Spearman range
None	0.048 ± 0.010	1.66	—
Lowest 10%	0.029 ± 0.007	1.27	0.935–0.966
Lowest 25%	0.019 ± 0.005	1.05	0.772–0.902

Propagation through preserved dynamics. The propagation experiments use three AIMove training seeds, four actions (SWML, MC_CLEAN, TVP_WRAP, and PLN_P02), three sampling seeds per action and model, and right forearm Z rotation row multipliers $0.25, 0.5, 0.75$: 36 sources and 108 intervention conditions. RMS trajectory differences use normalized channel coordinates over frames 6–120, after the five seed frames. Unedited operator rows, drive, and seed frames remain exactly unchanged. Cutting all contemporaneous and lagged connections from the target to the remaining channels eliminates their response in every condition while retaining 95.45–99.98% of the target response; each edited rollout is compared with its base under the same full or cut dynamics.

For the fixed-target ablations in Table 1, we hold the target difference trajectory and contemporaneous operators fixed and re-solve the remaining-channel difference dynamics. Each RMS ratio compares a separate ablation of lagged inputs or recurrence with the full dynamics, so the ratios do not form an additive decomposition. We average the three strengths within each source, then the 12 sources within each training seed, before computing across-seed statistics. These controls test transmission through the learned structural paths.

Interventions on different body regions. On the same Jacquard weaving motion, a treadle-leg intervention remains localized to the legs, while a trunk intervention reaches the arms, legs, and trunk (Figure A1). These patterns agree with the action-level coupling in Figure 3a.

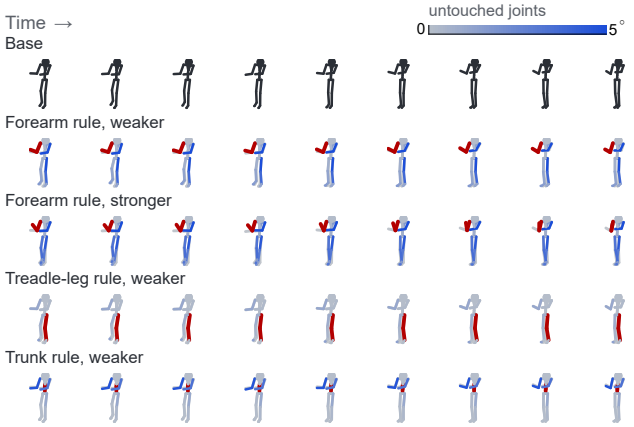

Figure A1: Responses to forearm, treadle-leg, and trunk row interventions on the same Jacquard weaving motion. Red marks the edited chain.

Value intervention decay. Figure A2 compares the realized effect of a value intervention with the decay bound of Proposition 1. Across the 25 evaluated actions, the largest solved gain over the sampled rollouts is 0.554710, close to the operational ceiling of 0.557. The observed deviation remains

below the corresponding per-sample bound and decreases substantially faster than its worst-case rate. The full-budget ceiling $g_\star \approx 0.985$ also covers amplification up to the admissible boundary.

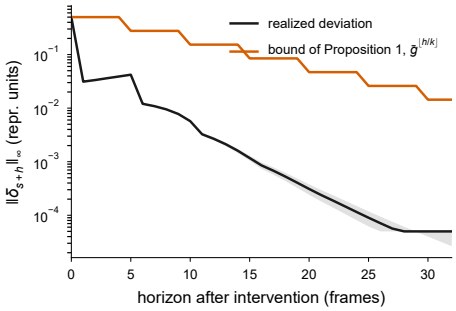

Figure A2: Value intervention response on AIMove at $k = 5$, shown against the per-sample bound of Proposition 1 for one training seed. Curves show the realized deviation and the band gives one standard deviation across the 25 actions. On the log scale, the response decays more quickly than the worst-case bound.

Additional mechanism interventions. Figure A3 extends the forearm row intervention of Section 4.2 to nine additional AIMove actions beyond Figure 4. Each pair retains the same latent trajectory, seed frames, and input drive while attenuating the right forearm Z rotation row by a factor of 0.5 across the contemporaneous and lagged operators.

A.3.2 Structural and Temporal Ablations

Structural ablations. Table 4 compares three variants under the same AIMove training recipe: 300 Stage 1 and 400 Stage 2 epochs, three independent training seeds, and checkpoint selection as specified in Appendix A.2.2. The reference uses the Stage 1 checkpoint with the lowest validation reconstruction loss and the final Stage 2 checkpoint. All variants use the generation metric protocol of Appendix A.2.2.

The variant without gain constraints tests the gain control construction as a whole by jointly removing the contemporaneous and lagged caps, their projections, and the stability penalty, while retaining DAG orientation. The variant without DAG orientation jointly changes edge orientation and final support density: it retains candidate thresholding and gain constraints but removes the ordering mask. The direct decoder replaces the structural solve with a framewise width-768 MLP, with approximately 691k decoder parameters versus 678k in MotionSCM. It retains the encoder and latent prior architectures but generates frames without autoregressive structural operators. Structural row interventions are consequently undefined for this control.

Requested and executed amplification. Attenuation uses the 0.5 whole-row edit and ROM readout of Section 4.2. For amplification, the incoming right forearm Z rotation row is multiplied by 2, 3, or 5 across contemporaneous and lagged operators. We compare the intervention interfaces under identical requests while allowing their executed changes to differ: constrained variants project requests onto the admissible operator set, while the unconstrained variant executes them directly. At each strength, the first eight prior draws per action are decoded anew at 360 frames, giving 200 rollouts per training seed. An explosion is a nonfinite normalized state or an absolute normalized state above 10.

In separate 180-frame probes, we quantify edit magnitude by the ratio of the executed to requested absolute coefficient *change*, complementing the 360-frame explosion tests. MotionSCM's ratios are approximately 0.50, 0.25, and 0.125 at $2\times$, $3\times$, and $5\times$, respectively; the unconstrained variant executes each request in full.

Complementing the uniform guarantee in Proposition 1, MotionSCM has no explosions at any tested amplification strength, with maximum absolute normalized state below 1.82 in the finite 360-frame tests. Without gain constraints, the explosion fraction rises from 0.038 ± 0.029 at $2\times$ to $0.472 \pm$

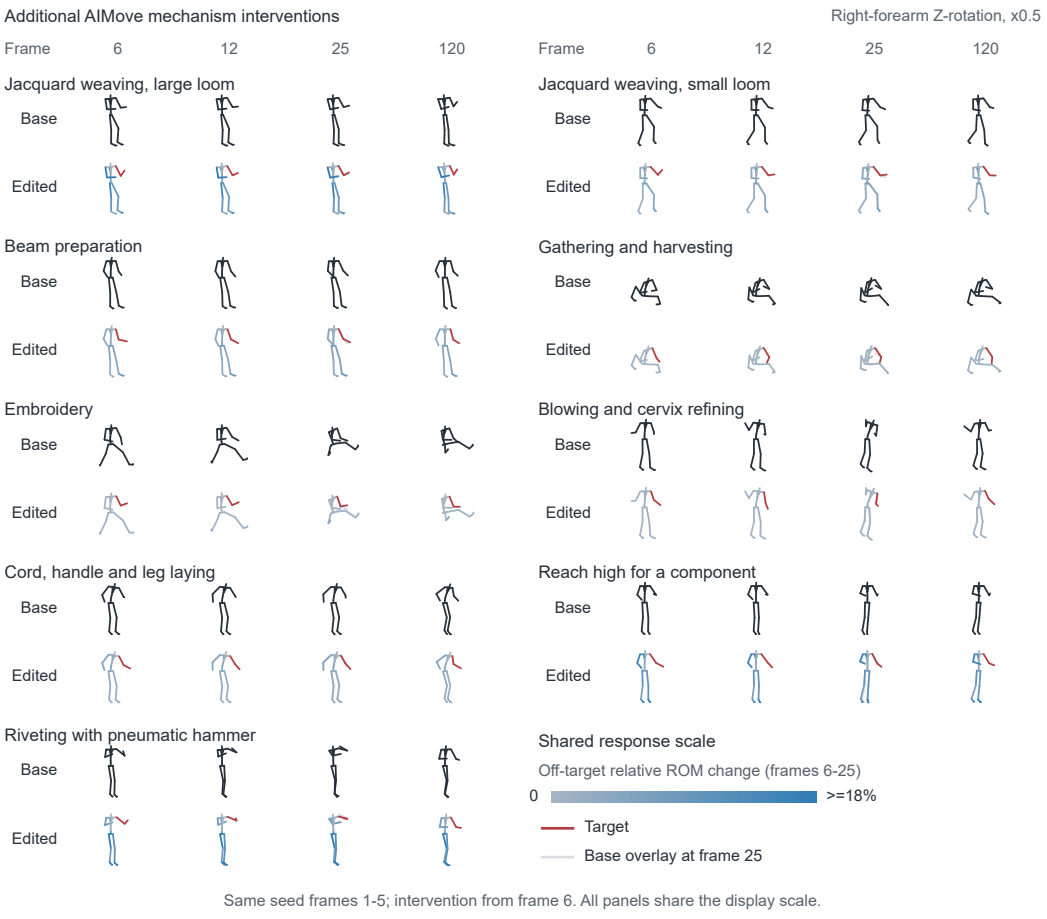

Figure A3: Mechanism interventions on nine additional AIMove actions. Each block pairs the base motion (top) with the edited motion (bottom) at frames 6, 12, 25, and 120. Red marks the target arm; blue encodes the relative ROM response over frames 6–25 on a shared scale capped at 18%. The base pose is ghosted in gray at frame 25.

Table A4: Coordination horizon ablation over three training seeds. Cross-joint response is the mean relative ROM response of the unedited channels; $\sigma_{\text{action}}/\sigma_{\text{seed}} > 1$ indicates action variation above seed variation.

k	FID ↓	Acc. ↑	Div. →	MModality →	Cross-joint response	$\sigma_{\text{action}}/\sigma_{\text{seed}}$
5	0.48±0.08	**0.89**±0.00	**6.44**±0.01	2.04±0.03	0.048±0.010	**1.66**
9	0.45±0.05	0.86±0.00	6.39±0.03	2.17±0.17	0.034±0.016	0.91
14	**0.38**±0.02	0.88±0.01	6.43±0.06	**2.14**±0.05	0.024±0.010	0.88

0.132 at 5×. Removing gain constraints also changes the lagged cross-joint coefficient share from 0.055 ± 0.005 to 0.706 ± 0.005. The maximum per-channel lagged gain, averaged over frames, changes from 0.427 to 3.257 ± 0.313. Thus removing the gain control construction changes the learned allocation as well as the response to amplification.

Coordination horizon. Increasing the coordination horizon does not substantially reduce the amount of learned coupling. Across three training seeds, the total coefficient mass of $\{\mathbf{A}_0, \ldots, \mathbf{A}_k\}$ is 52.1, 52.4, and 52.8 for $k = 5, 9, 14$, respectively, close to the aggregate operational cap $Dq(\tau_{\text{lag}} + \tau_{\text{con}}) = 53.5$. The corresponding cross-joint shares are also similar at 35.2%, 35.8%, and 34.8%. What changes with k is therefore not the overall amount of coupling, but how that coupling is distributed over time.

Figure A4 separates two aspects of this temporal distribution. We measure the absolute coefficient mass connecting the edited channel to unedited channels, separating contemporaneous and lagged contributions. Its contemporaneous contribution decreases as k increases. Across all lagged operators, the centroid of the total coefficient mass shifts from 34 to 72 ms.

These structural changes accompany the response variation in Table A4. Within each training run, the contemporaneous contribution of the learned path remains associated with which actions respond more strongly at all three horizons. At $k = 5$, the seed-averaged path contribution and cross-joint response have a Spearman correlation of 0.76. Separately, we measure the reproducibility of action-response rankings across training seeds: the mean Spearman correlation over the three seed pairs is 0.63, 0.43, and 0.37 for $k = 5, 9, 14$, respectively. Longer horizons retain cross-joint coordination, shift the global lagged mass towards longer delays, and yield less reproducible action-specific intervention responses across runs.

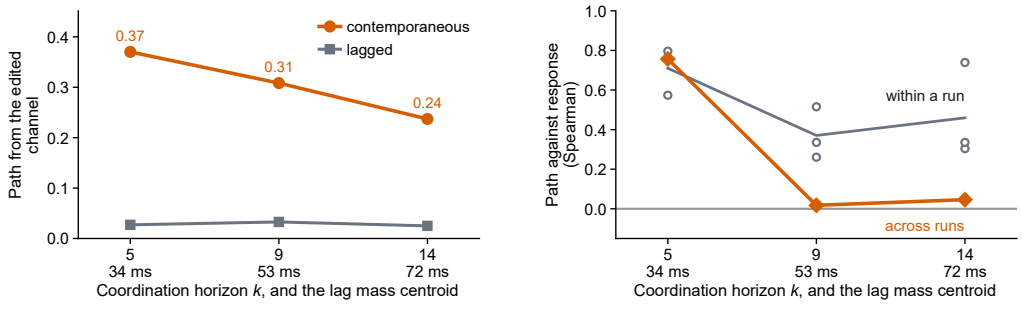

Figure A4: Effect of the coordination horizon over three training seeds. The left panel decomposes the path from the edited channel to the remaining channels into contemporaneous and lagged contributions; centroid labels summarize the total lagged coefficient mass across all channels. The right panel shows the rank correlation between this path's contemporaneous contribution and the per-action cross-joint response within each run. Open markers are individual runs, the grey line their mean, and filled markers the correlation from seed-averaged quantities.

Sensitivity to the budget constants. Using one training seed, we retrain the AIMove model under three perturbations of the reference setting that change the lagged and contemporaneous operating caps together. These use a tighter pair $(0.50, 0.30)$, a larger lag budget with $(0.80, 0.19)$, and a larger amplification reserve with $q = 1/2$. None of the tested perturbations degrades generation

quality relative to the matched reference run, and their overall FID variation is comparable to the variation of the reference configuration across its three training seeds. Action agreement also does not decrease under the perturbed settings.

The reference configuration produces the largest cross-joint response to the tested intervention. All three perturbations reduce the contemporaneous operating cap and the response relative to this reference. Table A5 shows that contemporaneous mass remains predominantly cross-joint and lagged mass predominantly within-channel across settings. Joint changes to the budgets alter this allocation: the tighter setting has the largest lagged cross-joint share, while its intervention response is the smallest. Relating coefficient shares to intervention responses therefore requires both absolute coupling strength and the dynamics that sustain effects.

Table A5: Coefficient allocation under the budget settings in Figure A5. Cross-joint shares are percentages of the respective operator mass; lagged utilization is the fraction of the operational per-channel budget used. We average four prior draws of 120 frames per action, with all 25 actions equally weighted and one training seed per setting.

Setting	Lagged cross-joint	Contemporaneous cross-joint	Lagged utilization
Reference	6.1%	93.0%	0.997
Tighter budgets	13.1%	91.7%	0.997
Larger lagged budget	5.1%	91.9%	0.998
Larger reserve	9.2%	94.5%	0.996

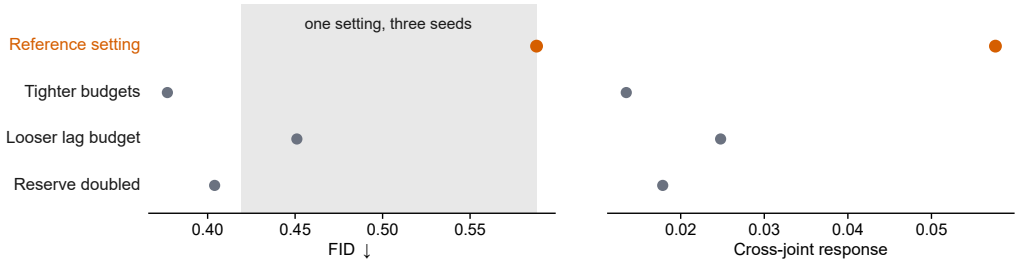

Figure A5: Sensitivity to the budget constants on AIMove. Each perturbed setting is retrained from one training seed. The generation panel also shows the variation of the reference configuration across its three training seeds, while the response panel reports the cross-joint response under the same intervention protocol.

A.3.3 Editor Comparison Details

Bone length consistency under matched editing. SALAD's five boxing sources attain target effects of 0.0184–0.0682 at the prescribed maximum attenuation (attention multiplier 0.125), below the matching interval $[0.09, 0.11]$ of Appendix A.2.2. They therefore have no matched pairs in either Table 2 or the geometric analysis below.

We evaluate all matched source–edit pairs in the three-action comparison of Table 2, using the 21 bones of the common 22-joint skeleton over frames 6–80. Let $\mathbf{L}^{\mathrm{b}}, \mathbf{L}^{\mathrm{e}} \in \mathbb{R}^{75\times 21}$ contain the base and edited bone lengths. We measure relative RMS bone length change and temporal coefficient of variation (CV) as

$$\Delta_L = \frac{\|\mathbf{L}^{\mathrm{e}} - \mathbf{L}^{\mathrm{b}}\|_{\mathrm{F}}}{\|\mathbf{L}^{\mathrm{b}}\|_{\mathrm{F}}}, \qquad \mathrm{CV}(\mathbf{L}) = \frac{1}{21}\sum_{b=1}^{21} \frac{\mathrm{std}_t(L_{t,b})}{\mathrm{mean}_t(L_{t,b})}.$$

Here $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. CV is evaluated separately for base and edited motion, using population standard deviation over frames. Table A6 reports the achieved target effects alongside the geometric readouts. These edits largely retain the bone length variability already present in their source motions. The off-target response ratio measures localization, while Δ_L and base/edited CV characterize preservation of source geometry.

Table A6: Achieved arm effects and bone length consistency for the three-action comparison. Entries are mean ± population standard deviation over five sampling seeds per method and action, following Appendix A.2.2. All entries use matched sources; SALAD boxing has no matches $(0/5)$.

Action	Method	Arm effect	Bone length change	Temporal CV (%)	
		r_S/a_S (%)	Δ_L (%)	Base	Edited
Walking	MotionSCM	9.784 ± 0.425	0.658 ± 0.090	6.709 ± 1.197	6.740 ± 1.188
	MotionCLR	9.905 ± 0.648	0.484 ± 0.102	1.837 ± 0.413	1.872 ± 0.390
	SALAD	9.878 ± 0.582	0.601 ± 0.066	3.986 ± 0.506	3.876 ± 0.462
Running	MotionSCM	10.169 ± 0.482	0.334 ± 0.077	4.625 ± 0.584	4.625 ± 0.586
	MotionCLR	9.728 ± 0.527	0.644 ± 0.319	2.431 ± 0.444	2.445 ± 0.393
	SALAD	9.868 ± 0.654	0.415 ± 0.056	4.397 ± 0.184	4.341 ± 0.185
Boxing	MotionSCM	10.273 ± 0.534	1.245 ± 0.411	6.961 ± 0.752	6.946 ± 0.745
	MotionCLR	10.187 ± 0.643	1.066 ± 0.296	4.278 ± 0.233	4.448 ± 0.187
	SALAD	—	—	—	—

A.3.4 Generation Diagnostics and Qualitative Results

Per-sequence speed. Figure A6 complements the generation results of Table 3 with an empirical comparison of corpus-level speeds, alongside the theoretical guarantee of rollout boundedness. For each sequence, we compute the 99.9th percentile of joint speed across all frames and joints in the recognizer's input space. We apply the same calculation to one sampling round per method and to the real corpus. On both position benchmarks, the median of this statistic for MotionSCM lies inside the 10th to 90th percentile range of the real corpus. The baselines miss that range in opposite directions, with EMDM developing an overly fast tail and MLD concentrating below it. On CMU, only 1.9% of MotionSCM samples exceed three times the real 99th-percentile speed, compared with 58.7% for EMDM.

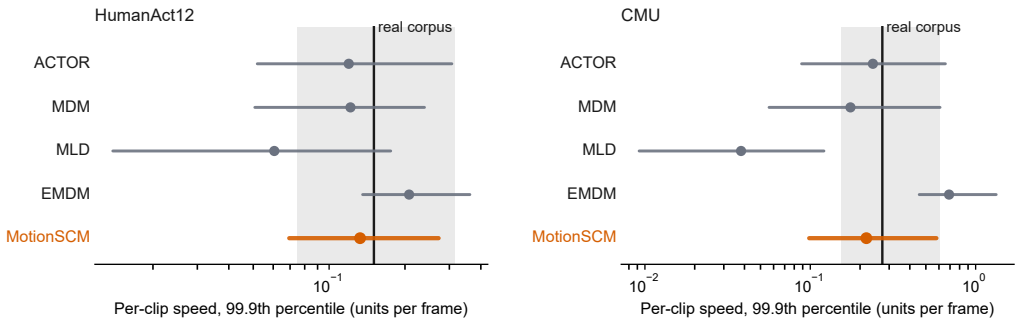

Figure A6: Per-sequence 99.9th-percentile joint speed in the recognizer's input space. Each bar spans a method's 10th to 90th percentile across sequences with its median marked; the shaded band and vertical line show the corresponding range and median of the real corpus.

Qualitative generation. Figures A7–A9 complement the generation results of Section 4.3 with examples from every action class. All model columns show independently generated motions. For each method, examples are selected by proximity to the real action class feature centroid within equal-sized candidate sets.

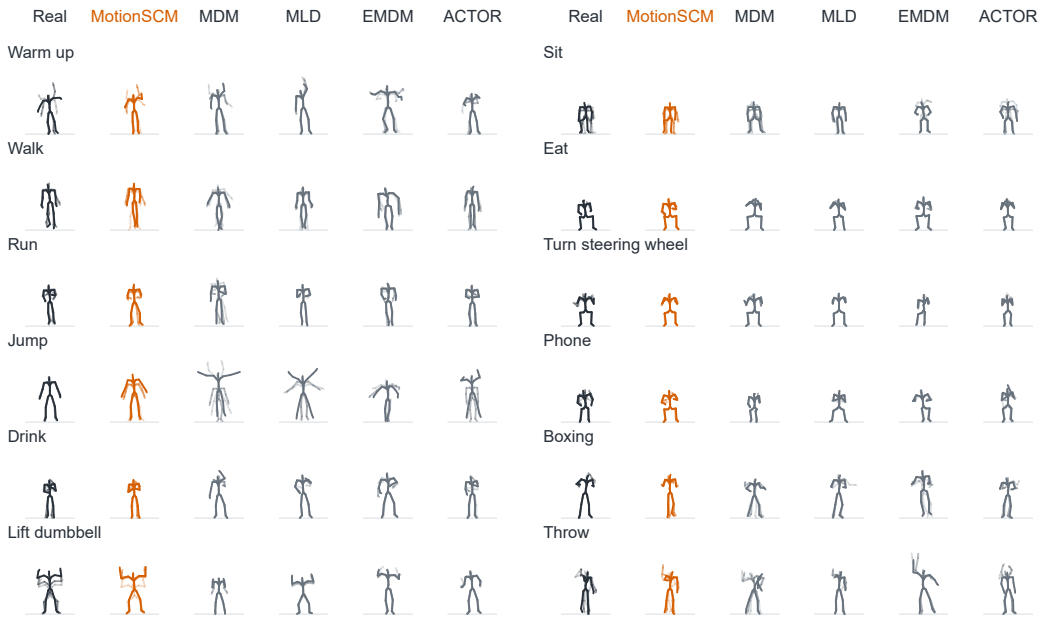

Figure A7: Qualitative generation on all 12 HumanAct12 actions. Each cell overlays poses at 30%, 62%, and 100% of its sequence, with darker poses occurring later. All columns share the rendering settings and skeleton scale.

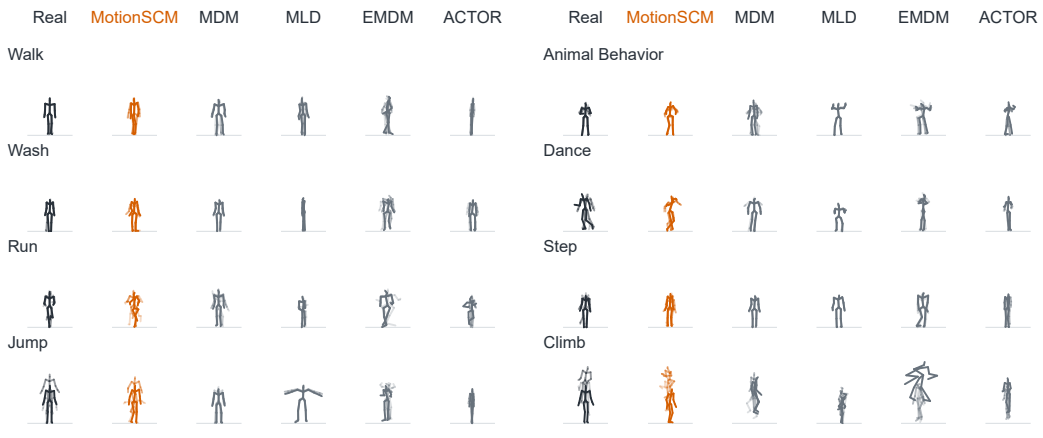

Figure A8: Qualitative generation on all eight CMU MoCap actions, with the same rendering conventions as Figure A7.



Figure A9: Qualitative generation on all 25 AIMove actions across six professional processes, with the same rendering conventions as Figure A7.