DANCEFIELDS: LEARNED DEFORMATION FIELDS FOR FEW-KEYPOSE DANCE STYLE TRANSFER

Anonymous authors
Paper under double-blind review

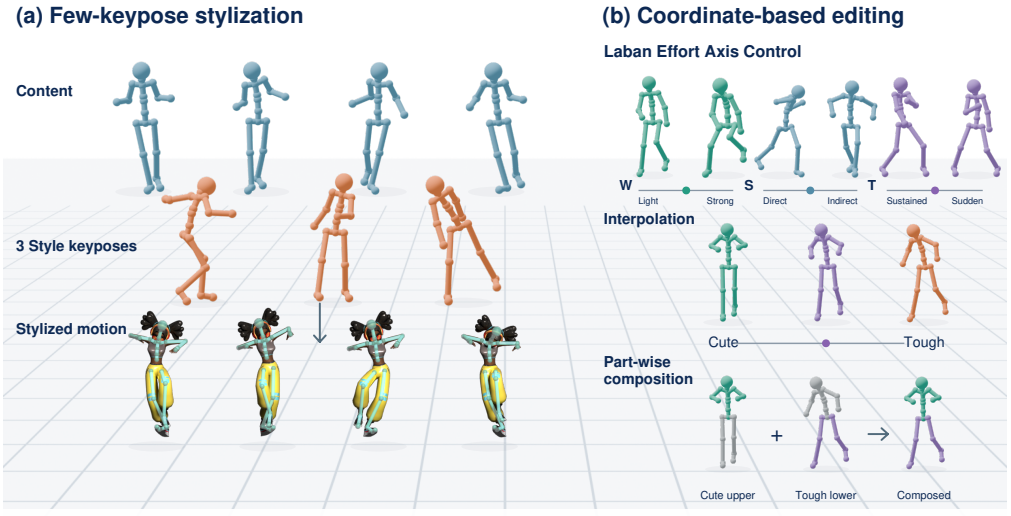

Figure 1: **DanceFields.** **(a)** From only three keyposes, DanceFields transfers an unseen style to an unseen choreography without per-style training. **(b)** Styles are represented as coordinates, enabling direct style editing.

ABSTRACT

Even for the same choreography, different dancers leave distinct impressions—style is not an accessory, but a core component that defines dance. While motion style transfer has predominantly focused on locomotion with consistent stylistic expressions across movements, dance fundamentally breaks this universality: the same style manifests through entirely different body parts and dynamics depending on the choreography. We formalize this choreography-dependent realization as the **reading** of a style and introduce the task of few-shot dance style transfer, which transfers expressive style while strictly preserving choreographic content. Because this artistic interpretation is extrinsic to visual pose features, it cannot be inferred from content alone and must instead be observed. We propose **DanceFields**, a representation that disentangles shared choreographic content from choreography-specific style readings. DanceFields learns a shared bank of content-conditioned deformation fields across choreographies, mapping a style to linear coordinates solved analytically in closed form via meta-learning. By observing merely three still keyposes of the target choreography in the desired style, our model resolves the style coordinates without per-style or test-time optimization. Furthermore, DanceFields explicitly models time as style, decoupling local temporal warping from global tempo variations. On a newly captured dataset of 30 choreographies across 16 styles (Laban Effort axes and character traits), DanceFields significantly outperforms per-style-trained baselines in style fidelity, content preservation, and physical plausibility. Crucially, representing style as coordinates naturally unlocks intuitive arithmetic controls—including continuous intensity grading, interpolation, and part-wise compositional editing—and demonstrates robust zero-shot generalization across unseen dance choreographies and even cross-domain locomotion benchmarks.

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.

1 INTRODUCTION

Even for the same choreography, different dancers leave distinct impressions—style is not an accessory, but a core component that defines dance. We refer to this property as **dance style** and define **dance style transfer** as transferring this style across performances while preserving the underlying choreography. Paired choreography–style data, in which the same choreography is captured both neutrally and in a style, is scarce, and capturing every choreography in every style is prohibitive in cost and labor; so *transferring to unseen choreographies from a few style examples* is the core of the task, not a convenience.

Existing approaches labeled as "dance style transfer" often transfer *genre* (Yin et al., 2023). However, genre dictates *what* movements are executed rather than *how* they are performed, functioning as content rather than style. In our formulation, aligned with (Goodman, 1976; Armelagos & Sirridge, 1978), **dance content** is the notatable order of keyposes, and **dance style** is what notation does not describe. Under this definition, timing and urgency inherently belong to style: traversing identical keyposes with varying speeds produces distinct stylistic impressions. We categorize style into two levels: physical dynamics via Laban's **Effort axes** (Weight, Space, Time) (Laban & Ullmann, 1971) and high-level **character** traits (e.g., cute, tough). Using 30 choreographies performed across 8 Effort actions and 8 character styles as paired data, we tackle this newly defined task. Three intrinsic properties distinguish our setting from locomotion style transfer, each motivating our technical solution as follows.

A deformation needs aligned motifs. Locomotion is summarized by a single action; dance is a sequence of motifs—a reach, a drop, a turn—each performed with its own timing. In locomotion, global or part-wise deformation therefore mattered (Jang et al., 2022; Aberman et al., 2020), while which moment is compared with which did not. In dance it does: *a difference taken between two performances at mismatched motifs is not a style difference*—it absorbs "the dancer got slightly further through the same choreography," which is content, not style. Phase representations provide such correspondence for periodic locomotion (Starke et al., 2022; Ji et al., 2025). Rather than a phase representation, we obtain it by aligning the two performances of a choreography onto a common progress axis—time-warping their joint trajectories when the pairs are built so that corresponding motifs coincide—and reuse that correspondence at every later stage. The deformation read at matched motifs is then what style is, and we represent it as *style coordinates* over shared *dance fields* (Sec. 3.2).

Alignment throws away rhythm, and rhythm is style. Urgency is exactly what notation omits, so it belongs to style, yet aligning the pair discards it. We therefore keep the time warp between the neutral and the styled performance—where the dancer rushed and where they lingered—as supervision for a separate time axis, and give timing its own *dance fields* and *style coordinates*: a bank of timing fields shared across choreographies, and per-performance coordinates over them that compose a log-rate profile (Sec. 3.3). Two things are read from that profile. *Local tempo* shapes a monotone time map, deciding where the performance rushes and where it lingers within a fixed sequence length—the shape of the timing, which is what characterizes Time (Sudden vs. Sustained) in Effort. *Global tempo* sets the overall pace of the performance, so the output is allowed to change length.

A style is a reading, and the reading changes. Locomotion style transfer deals with styles that are largely consistent across actions: *angry* manifests as thrust arms and kicked legs whether walking or running, allowing a single deformation to represent the style. Dance breaks this universality. For example, *cute* is expressed by drawing the upper body in for the lower-body dance *Butterfly*, but shifts to light foot hops in *HarlemShake* where upper-body space is occupied. This realization reflects an *artistic interpretation* that varies with the choreography, which we define as the **reading** of a style. Because this reading is extrinsic to the choreography's raw visual appearance, it cannot be inferred from poses alone (Sec. 4.2). What cannot be predicted must be observed: we analytically recover this reading from a few *style keyposes*—still poses of the target choreography performed in the style—and apply unseen styles without test-time training (Sec. 4.2). Conversely, styles with consistent readings across choreographies, such as the Effort axes, are applied *reference-free* directly via stored constants.

Representing style as coordinates makes this observation remarkably lightweight: rather than capturing an entire performance, we only need to identify a single point in a shared latent space. This formulation inherently enables versatile control—interpolation, signed and continuous intensity modulation, and part-wise mixing reduce to simple arithmetic within a unified coordinate system, providing the exact control flexibility required in practice. Because the reading is disentangled from the shared basis, an unseen style can be transferred to an unseen choreography from merely a few keyposes in just 2.5,ms. When evaluated zero-shot on the MOCHA dataset (Jang et al., 2023)—a locomotion benchmark whose rig, motions, and styles are entirely unseen during training—our model achieves the highest content preservation, delivers the best foot-contact score among all baselines, and ranks second in style metrics, demonstrating that cross-domain styles can be effectively inferred from minimal keyposes. Furthermore, on unseen AIST++ choreographies (Li et al., 2021), our geometrically defined Effort axes transfer completely reference-free without requiring any examples (Sec. 4.5).

Contributions.

- **The task.** We formulate *dance style transfer*, with content as the order of keyposes and style as kinesthetic motivation, and present, to our knowledge, the first attempt to address it.
- **Dance fields and style coordinates.** We propose a unified representation that disentangles shared choreographic content from choreography-specific style readings. Because styles are mapped into a shared coordinate space, versatile controls—including interpolation, part-wise composition, and continuous intensity scaling—naturally reduce to simple arithmetic operations.
- **Decoupled local and global temporal modeling.** We explicitly formalize time as a core component of style, separating it into local temporal warping and global tempo. This addresses two critical yet overlooked aspects in motion style transfer: variable output sequence durations and body-centric rhythmic execution beyond contact events.
- **Generalization.** Our formulation allows unseen styles to be applied to unseen choreographies in closed form and enables Effort axes to transfer reference-free. We demonstrate strong zero-shot generalization across unseen dance choreographies and even unaligned locomotion datasets.

2 Related Work

2.1 Dance style

In dance theory, style has long been understood through the distinction between *what* is danced and *how*. Phenomenological accounts locate the essence of dance in the *quality* with which a movement is performed—how force and time are carried in the body (Langer & Langer, 1953; Sheets-Johnstone, 2015)—and dance studies separate as style the layer that reads differently from performer to performer under a fixed choreography (Foster, 1986; Adshead-Lansdale & Layson, 2006); Laban gave that layer coordinates in Effort (Laban & Ullmann, 1971; Lange, 2012). Computational work on dance style, however, has stayed with the other half—what is danced. The only line of work to use the name "dance style transfer" transfers genre (Yin et al., 2023), and genre prescribes the movement vocabulary itself: changing it replaces the movements, so the choreography is not preserved and the transfer is one of content, not of style. **We take the distinction as dance theory states it**—hold the choreography, change only the quality with which it is performed—and read that quality on Laban's Effort axes and, above them, as character.

2.2 Motion style transfer

The task was formalized on walking and running (Xia et al., 2015; Holden et al., 2016): given a neutral motion X and a style indicator s (a label or a reference clip), the goal is to produce $Y = G(X, s)$. Methods differ mainly in *how s is injected*—AdaIN over feature statistics (Aberman et al., 2020), part-wise injection (Jang et al., 2022), graph latent codes (Park et al., 2021), style embeddings and guidance in diffusion models (Zhong et al., 2024; Wu et al., 2025; Hu et al., 2024;

Song et al., 2024), contact and trajectory factorization (Tang et al., 2024), and online constraints (Tao et al., 2022; Mason et al., 2022)—while MOCHA additionally retrieves a similar context, acknowledging that happiness in a jump is not happiness in a crawl (Jang et al., 2023). It treats the variation as a property of the context; we show in Sec. 4.2 that the reading of a dance style is not recoverable from the context under four similarity criteria, and observe it instead.

2.3 Few-shot and exemplar-based learning

Solving a new task from a few examples was formalized in image classification and language: MAML-style methods take gradient steps on the support set (Finn et al., 2017; Raghu et al., 2020), while R2-D2 and MetaOptNet replace that inner loop with a *closed-form base learner*, so what is learned is *a representation that solver can solve well* (Bertinetto et al., 2019; Lee et al., 2019). CAVIA isolates the task parameters into a few context variables (Zintgraf et al., 2019), a lineage continuing into in-context learning and task vectors (Akyürek et al., 2023; Hendel et al., 2023). In motion, a smaller set of methods targets low-data or single-example style learning. Li et al. (2022) synthesize new motions from a single sequence by modeling motion structure at multiple scales; Raab et al. (2024b) and Hu et al. (2024) adapt diffusion-based priors from a single style clip; Ponton et al. (2026) train a per-style model from a few seconds of a neutral–styled pair. A second group avoids per-style training altogether and takes a style clip as input at inference (Jang et al., 2022; Kim et al., 2024; Raab et al., 2024a). Our formulation sits between the two: the representation is trained once across choreographies, and a style is a small set of task parameters solved in closed form from a few observed poses of the target choreography.

3 Method

3.1 Dance style transfer, restated

Existing motion style transfer predicts the stylized motion directly from content and a style indicator, $Y = G(X, s)$, presuming that *the deformation indicated by s is well defined on X*. On locomotion it holds, which is why s can be stored in weights and applied to an arbitrary X; in dance it breaks, since the same *cute* is realized through different body parts depending on the choreography—the reading of Sec. 4.2, which changes with the choreography. That does not mean nothing is shared. Given a paired performance—a neutral content X and its styled performance Y, brought onto one time grid so that frame t of Y realizes frame t of X (Appendix A)—the *deformation* is the per-joint offset in root-relative positions, $D(t) = Y(t) - X(t) \in \mathbb{R}^{22\times 3}$. We require its model $\mathbf{\Delta}$ to split into two components:

$$\mathbf{\Delta}(t) \;=\; \mathbf{F}\big(c(t)\big)\,\mathbf{a}, \tag{1}$$

where $c(t)$ is the *content* at frame t, a fixed function of the neutral performance defined in Sec. 3.2 (Eq. 1 holds per joint, and we drop the joint index for brevity), $\mathbf{F}$ are the *dance fields*—shared across choreographies and styles, and learned—and $\mathbf{a}$ are the *style coordinates*—different per choreography and style, carrying the reading, and determined at inference.

3.2 Dance fields and style coordinates

Content ingredients. Joints $P \in \mathbb{R}^{T\times 22\times 3}$ are made root-relative and mapped to a 27-d feature: position (3), contrast against the temporal mean (3), velocity (3), acceleration (3), Gaussian band-pass residuals at three scales $\sigma = 2, 4, 8$ (9), a body-part one-hot (5) and an end-effector weight (1). *All of it is a deterministic function of P with no learned parameters*, and temporal–spatial context is folded inside, which is why the following MLP may be point-wise.

Dance fields. A two-layer GELU MLP f_θ maps the ingredients to $K = 64$ deformation fields $\mathbf{F} = f_\theta(c(t))$. It is applied to *one joint at one frame*, shared across all $22 \times T$ of them, and **does not take the style condition as input**. The fields are RMS-normalized to 1 over $(T, 22, 3)$, and a static mean column m_c—the time-constant mean deformation of condition c, computed from the training choreographies rather than predicted by f_θ—is appended as the $(K+1)$-th field. The learned fields therefore depend on the content alone: the same choreography yields the same first 64 fields under every style condition.

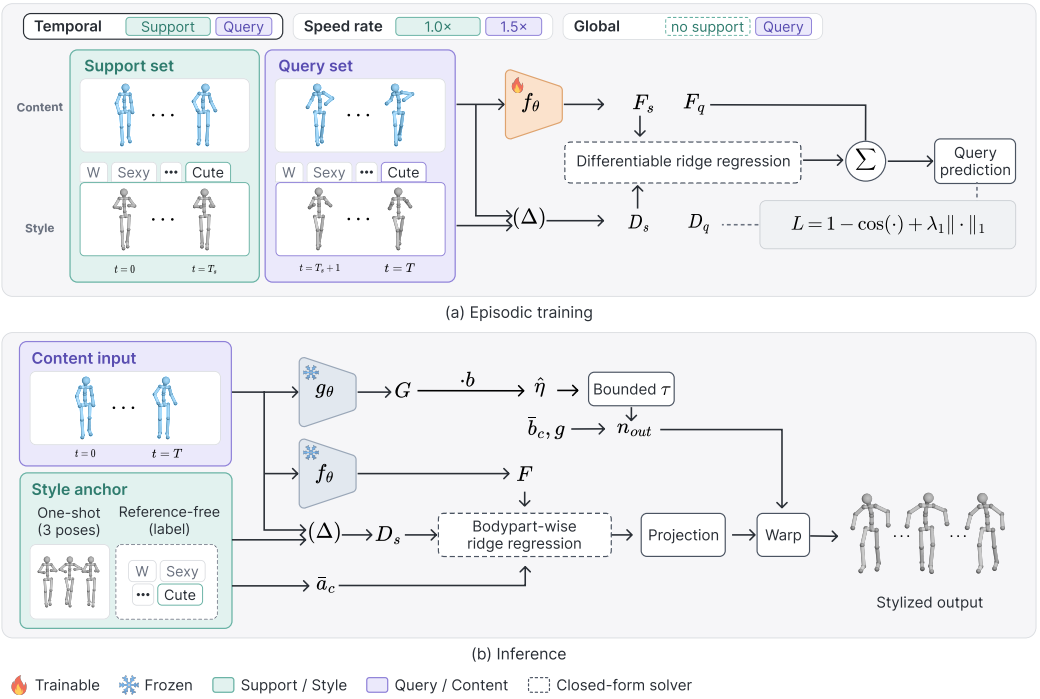

Figure 2: **Overview.** **(a)** Training is episodic: f_θ maps content alone to fields $\mathbf{F}$; the coordinates $\mathbf{a}_s$ are *solved* by a differentiable ridge on the support span and scored on the query span. **(b)** At inference f_θ, g_θ are frozen; $\mathbf{a}$ comes from a ridge on 3 keyposes centered on the stored constant $\bar{\mathbf{a}}_c$ (or $\bar{\mathbf{a}}_c$ alone), and $\bar{\mathbf{b}}_c$ drives the bounded time map. No per-style training.

Loss. The spatial branch combines a direction term and a magnitude term, $\mathcal{L}_{\text{sp}} = \text{mean}\big(1 - \cos(\hat{\boldsymbol{\Delta}}, D)\big) + \lambda_{L1}\|\hat{\boldsymbol{\Delta}} - D\|_1/\text{scale}$, with $\lambda_{L1} = 0.1$, where D is the ground-truth delta from the aligned pair. **There is no term that builds a basis**—no orthogonality, no diversity, no dictionary loss.

Training: closed-form meta-learning. Training follows the closed-form meta-learning structure of Sec. 2.3: a *support* set from which the coordinates are solved, and a *query* set on which the resulting deformation is scored. The unit is a 64-frame window of one aligned pair; a step draws a batch of eight windows and one episode type, which decides where the support comes from. `temporal` (probability 0.4, the deployment path) solves on the first 32 frames of the window and scores on the last 32, and vice versa; `speed rate` (0.3) solves on the same window played at another of eight playback speeds and scores at the original speed; `global` (0.3) has no support and reads a learned per-condition constant instead (discarded after training in favour of $\bar{\mathbf{a}}_c$ below). Each type mirrors a situation at inference: `temporal` reads the coordinates from part of a performance and applies them to the rest, as keyposes do; `speed rate` requires the same coordinates to hold when the performance is played at another speed, as it is once global tempo changes its length; and `global` predicts from a per-condition constant with no support, as the reference-free path does. The coordinates are solved in closed form. With $\mathbf{F}_s, \mathbf{F}_q = f_\theta$ on the two spans, D_s, D_q the aligned target deltas, both flattened over the frames and joints of the span, and $\ell = \mathcal{L}_{\text{sp}}$,

$$\begin{aligned} \mathbf{a}_s &= \text{ridge}(\mathbf{F}_s, D_s, \lambda) = (\mathbf{F}_s^\top \mathbf{F}_s + \lambda I)^{-1} \mathbf{F}_s^\top D_s, \\ \mathcal{L} &= \tfrac{1}{2}\big[\ell(\mathbf{F}_q \mathbf{a}_s, D_q) + \ell(\mathbf{F}_s \mathbf{a}_q, D_s)\big], \end{aligned} \tag{2}$$

with $\lambda = 0.01$: *we score where we did not read.* Because the ridge solution is closed-form, we backpropagate *through* it, so the gradient the MLP receives is not "fit the delta" but "produce fields a ridge can solve." Finally, `multi-cond` places four conditions of the *same* content in one batch; the fields are then identical and only the targets differ, so the coordinates are forced to carry the distinction—which they can only do if the fields spread far enough apart.

Why a basis emerges without a basis loss. Four wirings force it: the fields cannot receive the condition, so per-style fields are impossible; the coordinates are *solved, not predicted*, so a useless field gets coefficient 0 and therefore zero gradient, and only fields selected repeatedly across conditions survive; reading and scoring spans differ, so there is nowhere to memorize a per-choreography residual; and per-field RMS normalization blocks winning by growing, leaving competition in direction alone—which, as a side effect, makes coordinates comparable across choreographies, as the average below requires.

Style coordinates. The coordinates are not learned. They arrive by two routes.

(a) Observation—prior-centered part-wise ridge. From k still poses of the target choreography in that style ($k=3$ at deployment), for each of 5 body parts:

$$\mathbf{a} \;=\; \bar{\mathbf{a}}_c \;+\; \mathrm{ridge}\big(\mathbf{F}_s,\; D_s - \mathbf{F}_s\,\bar{\mathbf{a}}_c,\; \lambda_p\big), \qquad \lambda_p = 30, \tag{3}$$

where the support set s is now the k anchor poses. Centering the regularizer on the style constant $\bar{\mathbf{a}}_c$ rather than on 0 is decisive: at 0 the anchor *replaces* the constant, at $\bar{\mathbf{a}}_c$ it *corrects* it (character direction cosine with one pose: 0.569 vs. 0.333, below even the reference-free 0.450; Table 6). The choice of λ_{ap} is discussed in Appendix F.

(b) The constant $\bar{\mathbf{a}}_c$—an average back-solved from training choreographies. It is computed once after training: for each of the fold's 21 training choreographies, we back-solve the optimal coordinates against the GT delta, average them, and apply a magnitude correction

$$\bar{\mathbf{a}}_c \;=\; \bar{\mathbf{a}}_c^{\mathrm{raw}}\,\frac{\sum_d \mathrm{RMS}(D_d)}{\sum_d \mathrm{RMS}(\mathbf{F}_d\,\bar{\mathbf{a}}_c^{\mathrm{raw}})}, \tag{4}$$

needed because averaging directions cancels magnitude—the Weight deformation shrinks from 3.68 to 1.18 cm—in a way the scale-invariant cosine does not see. This average also beats a *learned* per-condition constant head (reference-free character cosine 0.430 vs. 0.413), and agrees at cosine 0.986 with a constant tuned to maximize the evaluation metric directly, which itself *drops* on test (Table 7).

Finally $\boldsymbol{\Delta} = \mathbf{F}\mathbf{a} = \sum_k \mathbf{F}_k a_k$ is added to the content, passed through a deterministic bone-length and contact projection, and warped by the time map of Sec. 3.3. **Everything after the fields has no learned parameters** which is why adding a style costs zero gradient updates. The fields themselves are learned; the precise statement is *no per-style training*, not *no training*.

3.3 Local tempo and global tempo

We apply the same structure (Eq. 1) to the separated time axis, with two changes: the fields are scalar, and the output is a time map rather than a displacement. A two-layer MLP g_θ maps an 83-d timing input—the joint ingredients of Sec. 3.2 averaged over joints, plus 56 speed-, acceleration- and jerk-based features of the trajectory (Appendix B)—to $K_t = 12$ fields $\mathbf{G} = [g_k(t)]$, and timing coordinates $\mathbf{b}$ compose a log-rate profile

$$\hat{\eta} = \mathbf{G}\mathbf{b}, \qquad \mathrm{inc} = \mathrm{clamp}_{[0.35,\,2]}\big(\exp(\pm\kappa\,\hat{\eta})\big), \qquad \tau = \mathrm{cumsum}\big(\mathrm{inc}/\textstyle\sum \mathrm{inc}\big), \tag{5}$$

where $\mathrm{inc}[i]$ is the playback rate of interval i, $\pm$ the pole of the axis and κ its strength ($\kappa = 2$ at deployment). *Local tempo* is this map τ: where the performance rushes and where it lingers, which is what separates Sudden from Sustained. *Global tempo* sets the output length, $n_{\mathrm{out}} = n/g^{\pm\kappa}$, from a per-condition speed g (1.053 over training choreographies), so the performance may become faster or slower as a whole. $\mathbf{b}$ is solved by ridge in training exactly as $\mathbf{a}$ is; at deployment it is the stored constant $\bar{\mathbf{b}}_c$, since still keyposes carry no timing. The time loss adds $\lambda_{\mathrm{amp}}(\log \mathrm{RMS}(\hat{\eta}) - \log \mathrm{RMS}(\eta))^2$ to the direction and L_1 terms: cosine is blind to how strongly the performance speeds up and slows down, and without this term $\hat{\eta}$ keeps its direction but shrinks in magnitude—the same loss of magnitude that the averaged constant suffers in space, corrected there at inference and here in the loss.

4 Experiments

4.1 Setup

Data and folds. An in-house capture of 30 choreographies $\times$ 16 styles (the 8 Basic Effort actions, which give the two poles of Weight, Space and Time, and 8 character styles), each performance a

styled rendition of the same neutral take, split into six choreography-disjoint folds of 21 train / 4 validation / 5 strict-test choreographies. Test ground truth is used for scoring only; the mean column m_c and the constants $\bar{\mathbf{a}}_c$, $\bar{\mathbf{b}}_c$ are recomputed per fold from its training choreographies.

Metrics and tests. The direction cosine is taken between flattened root-relative joint residuals over the whole clip (P-full). CC/SC (Kim et al., 2024) are mean per-frame L2 (cm) between output and content, and between output and the ground-truth performance of the same choreography in the target style, in a common canonical frame. PFC (Tseng et al., 2023) scores physical foot contact without any contact label: the mean over frames of the upward-clipped centre-of-mass acceleration times the speeds of both feet, normalized by the peak acceleration ($\times 10^{-3}$, lower is better). ΔJerk follows STyMo (Ponton et al., 2026): the change in mean joint jerk relative to the source (%). Roots are taken from the source for every method except STyMo, which supplies its own. Differences are tested by paired t-tests over folds ($n = 6$, Holm-corrected within a table) or per choreography for baseline SC ($n = 29$), and intervals are choreography-clustered bootstrap 95% CIs.

Baselines. STyMo requires pairs with identical filenames and frame counts, so it received pairs resampled by our DTW alignment with headers re-stamped at the measured frame rate; SinMDM and GANimator were trained per style on `BobbyBrown`, which is therefore excluded from evaluation for all methods; MoST and MoMo use released checkpoints. For ours, *reference-free* means no frame of the target choreography is seen; the constant $\bar{\mathbf{a}}_c$ still comes from the training choreographies, so we write *no per-style training*, not *no training*.

4.2 The reading of a style cannot be predicted, only observed

Within one style, how the deformation is realized changes from choreography to choreography (Fig. 4, appendix: *Tough* tilts the torso in one, drops the weight into the legs in another, barely reaches the arms in a third). Solved per choreography, the coordinates of a single character style point in nearly orthogonal directions (across-choreography cosine 0.074; Table 3, appendix). This component is not written in the choreography. Under four notions of choreographic similarity—mean pose, range of motion, velocity profile, and the field network's own input features—similarity explains under 2% of the variance in directional agreement, so no retrieval rule can find the right coordinates.

Observation needs only a few scattered poses. What does recover the reading is observation. Three still poses of the target choreography close most of the gap between the reference-free constant and the full-clip oracle (character direction cosine $0.448 \rightarrow 0.683$, oracle 0.784; Table 2, appendix). That gap is the size of the component, and it is what our representation keeps in the coordinates and out of the weights. For authoring, what matters is that *three poses scattered through the performance beat a whole contiguous second of it* (0.683 vs. 0.647): neighbouring frames are near-duplicates to the solve, so coverage of the performance, not duration of footage, is what provides information. No selection rule we tried reliably beats uniform sampling. The reading tolerates centimetres of joint noise but not a few frames of phase error: a 5-frame offset drops the Effort cosine from 0.607 to 0.222, below reference-free (Table 7, appendix), so keyposes must be placed at their true moment on the progress axis (Appendix A).

4.3 Style is transferred and perceived

The study asks whether the transferred style is *perceived* (Fig. 3). For Effort, participants learn a polarity from one ground-truth pair and find it again in our output on another choreography. With *no example given to the model*, our Effort axes are read as accurately as real performances (statistically equivalent to the human ceiling), including Space (74% against a 75% ceiling): a fixed reading transfers even without a fixed direction. Character styles cannot be evaluated the same way: even on real performances, the eight labels are not reliably told apart, so we do not ask participants to name one. Instead, we split the styles into two groups of similar character, and in each trial show one real performance of the target style on a *different* choreography together with three candidates on the test choreography—our output for the target style and two outputs for styles from the other group—asking which candidate moved toward the target. Participants pick the right one at 92% of the rate they do when the candidates are real performances; the grouping and full protocol are

in Appendix E. Against each baseline separately, participants preferred our output for both style proximity and naturalness: 93%/93% over MoST and 79%/86% over STyMo.

Table 3 confirms the split of Sec. 1 on ground-truth deltas: the deformation direction of Weight is fixed across choreographies (split-half cosine 0.88), whereas those of character styles and Space change (0.07 and 0.02). Space, however, is an Effort axis whose reading is fixed even though its direction is not: "more direct" means the same thing in every choreography. Whether such a reading transfers is a question of perception rather than geometry, so we answer it with a user study.

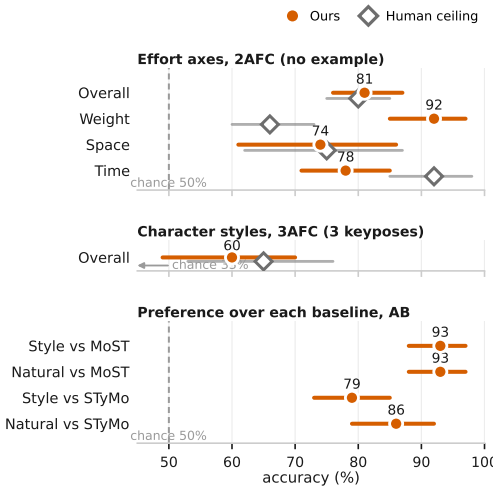

Figure 3: **User study.** Dots are the model, open diamonds the human ceiling on real performances, dashed lines chance; bars are 95% CIs. On axis discrimination the model is given no example and is equivalent to the ceiling.

4.4 Beating per-style-trained baselines without per-style training

Fig.5 (appendix) shows the comparison before any metric: MoST drifts from the target root and STyMo loses the pose, while ours stays on the target on an Effort axis and a character style alike. Table1 quantifies it. With no example of the style, ours already has the lowest CC and SC of any method on the Effort axes; with three keyposes it is closest to the ground truth on both families (SC 40.2 / 48.6), and with five it is at once the least damage to the content and the shortest distance to the target (CC 31.0 / SC 37.2). Against every baseline the three-keypose SC difference holds on 29/29 choreographies (Holm $p \leq 1.5 \times 10^{-5}$). On physical plausibility, PFC ranks ours first on the Effort axes (0.25 vs. STyMo's 0.34) and STyMo first on character styles (0.41 vs. 0.67), where character deformations reach the legs and move feet that the neutral take kept planted. STyMo avoids this by barely moving them, and pays for it in ΔJerk: at -444 / -683 its output is far smoother than the source, whereas ours stays within the ground truth's own offset. A new style here is a closed-form solve of a few milliseconds that adds no parameters; the per-style baselines multiply their model count by the number of styles—STyMo alone, trained per style and per fold, cost roughly nine hundred GPU-hours for this table, against none for ours.

Table 1: Comparison to style-transfer baselines. Best in bold, second underlined; Identity is the no-deformation reference. Confidence intervals in Appendix G.

		Effort				Character			
Method	What it received	CC↓	SC↓	PFC↓	ΔJerk→0	CC↓	SC↓	PFC↓	ΔJerk→0
GANimator (Li et al., 2022)	per-style	83.4	85.8	0.91	+114	80.4	83.7	1.50	**+56**
SinMDM (Raab et al., 2024b)	per-style	106.0	105.5	1.26	+183	96.6	95.5	1.80	+121
MoST (Kim et al., 2024)	released ckpt	70.6	78.7	0.93	+318	70.8	79.4	1.57	+256
MoMo (Raab et al., 2024a)	released ckpt	87.3	91.2	3.30	+1879	88.6	89.8	5.04	+2282
STyMo (Ponton et al., 2026)	per-style, 21 pairs	54.4	63.0	0.34	−444	60.4	71.1	**0.41**	−683
Ours	none, **0 examples**	42.9	52.3	0.30	**+14**	65.6	69.2	0.99	+282
Ours	none, 1 keypose	**30.5**	43.1	0.26	+106	**44.1**	53.9	0.67	+105
Ours	none, 3 keyposes	32.1	40.2	**0.25**	+120	46.5	48.6	0.67	+133
Ours	none, 5 keyposes	31.0	**37.2**	0.26	+135	46.3	**44.8**	0.71	+146
Identity	*no deformation*	*0.0*	*44.5*	*0.19*	*0*	*0.0*	*59.5*	*0.45*	*0*

4.5 Generalization and editing

Without retraining, the fields also transfer to the 25-joint MOCHA rig, its eight locomotion styles and its evaluation content, none of which appear in training. With the style read from a single paired clip, ours preserves content best and has the best physical foot-contact score among all methods, including models trained on that clip (Table 11, Appendix H). On AIST++ the transfer is tested in two ways (Fig. 8, appendix). Some AIST++ choreographies coincide with ours but are danced by other dancers in another capture; on these, character styles read from three keyposes and reference-free Effort axes land as they do on our own data, so the fields depend on neither the dancer nor the capture. And although the fields are trained on hip-hop only, the Effort axes transfer to, and remain controllable on, choreographies from genres that never appear in training, such as ballet, house and breaking.

Because a style is a point in one coordinate system, the controls a choreographer wants are arithmetic on that point, with no retraining between settings: moving along the segment between two coordinates interpolates between the styles, scaling a coordinate grades its intensity with sign (Fig. 6, appendix), and, since coordinates are solved per body part, the upper body can take one style and the lower body another (Fig. 7, appendix).

4.6 Ablations: what is lost when a component is removed

Tables 5–7 (appendix) remove one component at a time and identify what each one carries. Without *learned* fields, character transfer collapses ($0.594 \rightarrow 0.267$); fixing one coordinate per style collapses the per-choreography reading to a single direction ($0.683 \rightarrow 0.448$), and that gap *is* the changing reading component; centering the ridge on 0 instead of the prior lets one anchor replace the constant and fall below reference-free ($0.569 \rightarrow 0.333$); and a 5-frame alignment offset drops the Effort cosine below reference-free ($0.607 \rightarrow 0.222$), which re-alignment restores (0.596).

5 Conclusion and Discussion

We have defined the problem of transferring dance style while preserving the choreography. In dance, the same choreography, and even the same movement motif, can be expressed in markedly different ways, and this variation in interpretation is the core of dance style. Because such interpretation cannot be inferred from the choreography by learning alone, DanceFields decomposes the style deformation into a component that is shared and a component that is not, learning the former from data while keeping the latter specific to each interpretation. This decomposition allows a deformation field learned from only a few examples to generalize to unseen choreographies while still transferring the distinctive character of the target style, yielding strong results in both generalization and transfer quality.

Limitations and future work. Our approach has two main limitations. First, the style interpretation is fixed over a sequence and does not change over time, so the model cannot capture a performance whose character evolves, for example a dancer who gradually intensifies toward a climax. Second, our method relies on accurate correspondence between the paired performances: the timing of each keypose is inherited from the pair alignment described in Appendix A, and errors in this alignment propagate directly to the transferred style, as shown by the alignment ablation in Table 7. Making the method robust to imperfect alignment, for instance by jointly refining correspondence and style or by learning the alignment end-to-end, and modeling time-varying interpretation are natural directions for future work.

Ethics statement

The dance motion capture used in this work was collected with informed consent from the performers for research use, and the user study was conducted with informed consent from all participants. Because the performers consented to research use only, we do not release the captured data.

REPRODUCIBILITY STATEMENT

Sec. 4.1 specifies the folds, metrics, identity floors and statistical tests; Sec. 3 specifies features, architecture, losses, episode sampling and both closed-form solves with their hyperparameters. The appendix gives the user-study protocol, the full ablation table and the cross-rig comparison.

AI USE STATEMENT

Large language models were used for editing assistance and for literature search; all experiments, measurements and claims are the authors'.

A ALIGNMENT: THE PRECONDITION FOR A DEFORMATION TO EXIST

A deformation is defined only between matched phases; mapping frame indices directly mixes in "the dancer got further through the same choreography." Our pairs are returned to a shared grid when they are built and that correspondence is *frozen* downstream. Splitting it into two warps is essential: the **pair warp** between neutral and styled supervises the time axis, the **mapping warp** aligns for spatial comparison. Merging them—reading timing back out of already-aligned motion—does not recover local timing (cosine 0.012 with the true η; Table 7).

How the warps are computed. Both correspondences come from a single alignment made when a pair is built. Each take is *constant-speed resampled*: its summed root-relative joint displacement is accumulated into an arc-length progress and the clip is resampled uniformly in it onto $N = 160$ points, so that the axis is "how much has been danced" rather than time. The two constant-speed curves are then aligned by dynamic time warping on root-relative joint positions with the facing rotation removed, so that a difference in heading is not mistaken for a different pose; no per-clip statistics enter the features, since cropping would change them. The alignment is local with free start and end: within a band of 35% of the clip, each step earns the margin by which its pose distance beats a threshold set at the 35th percentile of in-band distances, and the path of maximum accumulated margin is taken, which extends the overlap for as long as the poses genuinely correspond. The neutral take's grid is the canonical progress axis of the choreography and the **mapping warp**: every styled take is carried back onto it by inverting its alignment path, and the spatial deformation D is always read there, on the canonical 22-joint skeleton. The same path, read as a map from neutral time to styled time, is the **pair warp** ϕ; the timing target is its log-rate, $\eta = \log \phi'$, taken from progress curves smoothed with $\sigma = 2$ frames and centred at its median. The grid is then frozen: the training set is built on the intersection of the styles' valid ranges, no second alignment is ever run, and keyposes supplied at inference are placed on the same axis by their progress index—which is why an offset of a few frames (Table 7) is enough to misattribute the deformation.

B TIMING FEATURES

The 56 timing features are deterministic functions of the root-relative trajectory, each z-scored within the clip: whole-body speed, acceleration and jerk (norms summed over joints), the summed magnitude of vertical acceleration, a slow-moment emphasis $\exp\big(-(\text{speed} - \min \text{speed})/\sigma_{\text{speed}}\big)$ that peaks on beats, speed and acceleration smoothed with Gaussians of $\sigma = 2, 4, 8$ frames, tangential acceleration (the forward difference of speed), and per-joint speed and acceleration magnitudes (22 each). Training and evaluation windows for the time axis skip the first 16 frames of each clip, where the ground-truth η carries an alignment-boundary spike ($|\eta| = 0.607$ on frames 0–7 vs. 0.134 mid-clip, averaged over the training windows of all 30 choreographies).

C KEYPOSE COUNT AND AXIS STRUCTURE

Table 2 is the measurement behind Sec. 4.2: how far the style direction cosine rises as the observation grows from nothing (the stored constant) to a few still keyposes to the whole target clip. Three poses scattered through the performance beat a whole contiguous second of it, because neighbouring frames are near-duplicates to the solve. Table 3 is the measurement behind Sec. 4.3: on ground-truth

deltas without any model, whether a style's deformation keeps a shared direction across choreographies. Weight does (split-half cosine 0.88, across-choreography 0.38), whereas character styles (0.07) and Space (0.19 / 0.02) do not.

Table 2: Style direction cosine vs. the number of observed style keyposes (P-full, six-fold).

Anchors	Character	W	S
none (reference-free)	0.448	0.406	0.042
1 pose	0.578	0.528	0.266
3 poses	0.683	0.607	0.427
5 poses	**0.730**	**0.690**	**0.541**
contiguous 32 frames (1.07 s)	0.647	—	—
full clip (oracle)	0.784	0.748	0.634

Table 3: Whether a style's deformation keeps a shared direction across choreographies, measured on GT deltas without the model.

Axis	Split-half	Across-choreo.	Ref.-free
Weight	0.881	0.376	0.406
Character	—	0.074	0.448
Space	0.193	0.015	0.042

D QUALITATIVE RESULTS

One style, different readings (Fig. 4). The same character style applied to three choreographies lands on different body parts: *Tough* tilts the torso in JanetJackson, drops the weight into the legs in SteveMartin and barely leaves the arms in Bernie, and *Cute* and *Comic* shift in the same way. Reading down a column, what survives across choreographies is the character; reading across a row, how it is realized is decided by the choreography. This is the reading of Sec. 4.2 seen directly, and it is why the coordinates are read from keyposes of the target choreography rather than stored per style.

Against the baselines (Fig. 5). On both an Effort axis (Bernie $\times$ Wring) and a character style (JanetJackson $\times$ Tough), ours stays on the styled ground truth through all four phases, so the coloured skeleton overlaps the grey one. STyMo keeps the root but the pose breaks in places: limbs fold or overshoot where the target does not, which is the motion damage that its low ΔJerk in Table 1 hides. MoST does not follow the content at all: its root drifts from the ground truth and the poses belong to a different movement, which is what its CC of 70 cm measures. Ours is given three keyposes; both baselines are trained per style.

Signed and graded Effort control (Fig. 6). Scaling one coordinate scales the deformation. Reading each row from $-3\times$ to $+3\times$, the offset between the orange output and the grey neutral grows with the scale and reverses with its sign, so the two poles of Weight, Space and Time are reached from one coordinate; along Space the offset is smallest, consistent with Space being the weakest axis on ground truth (Table 3). The sign reverses the direction and the scale grades it, with no retraining between cells: the coordinates are simply multiplied.

Editing in coordinate space (Fig. 7). Because a style is a point, the segment between two points is a style: moving from *Tough* to *Cute* on two unseen choreographies changes the output continuously, with every intermediate pose a valid one. And because the coordinates are solved per body part, the parts can be edited separately: in the bottom rows the arms take one of two styles (*Chessy* or *Swag*) and the legs the other (*Tough*), and swapping the assignment swaps the parts, so a choreographer can change how the legs are carried without touching the arms.

E USER STUDY PROTOCOL

25 participants, 1,275 valid trials. Three blocks: (i) *axis discrimination*, a 2AFC per Effort axis in which the participant first sees one ground-truth $\pm$ pair of a choreography that appears in no trial, then decides which of two model outputs on other choreographies is the $+$ pole; (ii) *character identification*, a 3AFC: the reference is a real performance of the target style on a choreography that appears in no candidate (*ToneWop*); the three candidates are, on one test choreography, our output for the target style and two outputs for styles from the opposite group of the split *Tough*, *Comic* vs. *Cute*, *Sexy*, *Swag*, *Fresh*, *Chessy*. The split is the one on which two ground-truth descriptors (posture direction and limb dynamics) agree; *Clumsy* changes side between them and is excluded. The ceiling

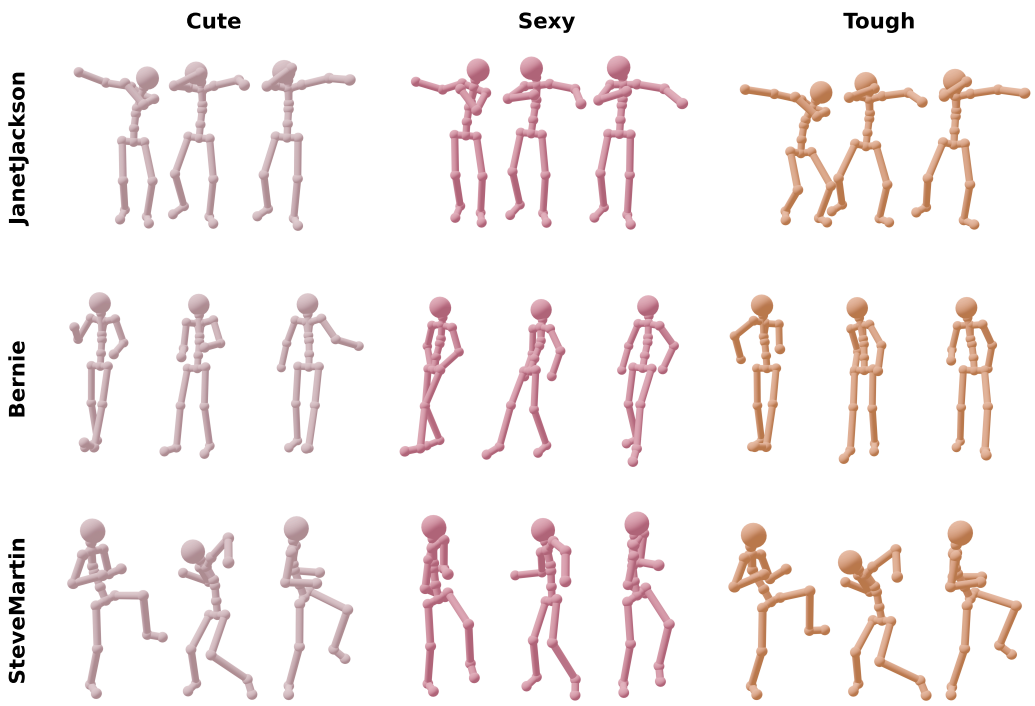

Figure 4: **One style is not one deformation.** The same three character styles applied to three choreographies, three poses each. Reading down a column shows what survives across choreographies; reading across a row shows how much of the realization is specific to the choreography. *Tough* tilts the torso in JanetJackson, drops the weight into the legs in SteveMartin and barely leaves the arms in Bernie. This is the reading, which cannot be predicted from content and has to be observed (Sec. 4.2).

condition replaces the candidates with real performances; (iii) *preference* against MoST and STyMo for style proximity and naturalness. The human ceiling is the same task on real performances. Per-axis accuracies are Weight 92%, Space 74%, Time 78% against ceilings of 92%, 75% and 92%; the 2AFC block is statistically equivalent to its ceiling (TOST $\pm 10\%$p, $p = 0.0024$); the 3AFC block reaches 60% against a 65% ceiling; preference is 86% (style) and 89% (naturalness) pooled; per baseline, 93%/93% against MoST ($n = 124$) and 79%/86% against STyMo ($n = 126$), all $p < 10^{-8}$. Weight exceeds its ceiling because our deformation is $1.13\times$ the ground-truth amplitude at the shown strength, not because the model beats people. The preference block used a pre-final stimulus for the four Effort styles and is a conservative lower bound. Intervals are participant-clustered bootstrap 95% CIs.

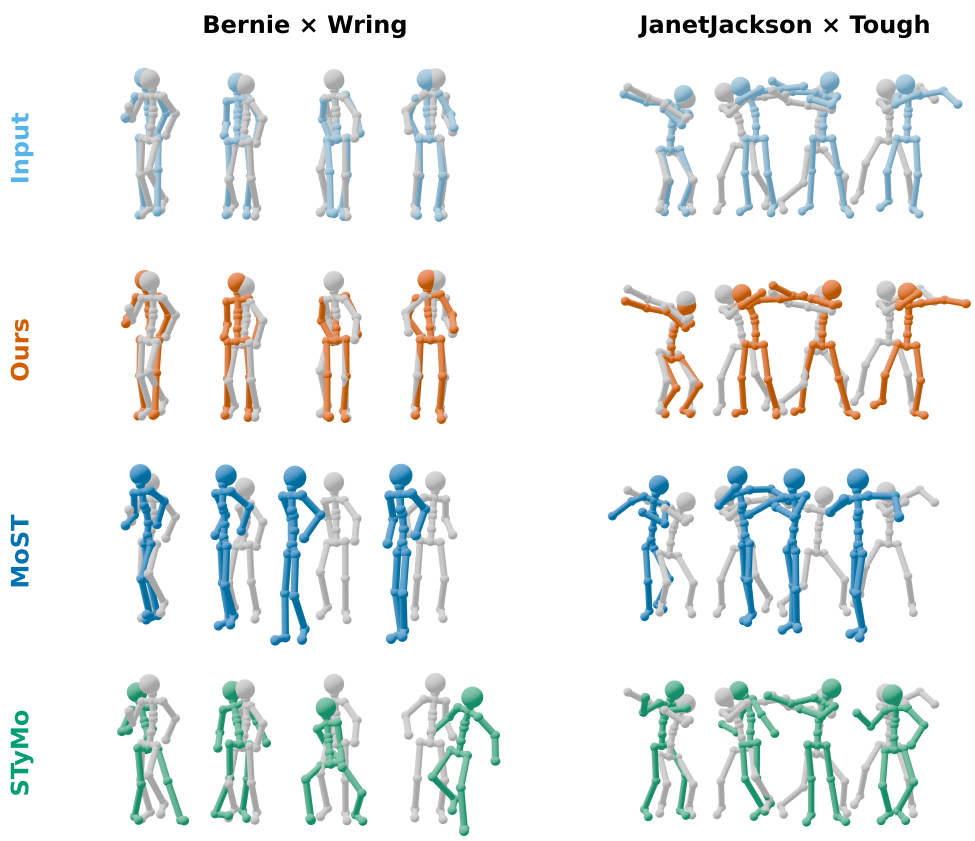

Figure 5: **Qualitative comparison on unseen choreography.** Four phases of one performance; grey is the styled ground truth, colour is each method's output, and the distance between them is the error. The *Input* row shows how far the style has to move the neutral take. Ours follows the target on both an Effort axis (Bernie × Wring) and a character style (JanetJackson × Tough), while MoST drifts away from the ground-truth root and STyMo loses the pose. Ours is given three keyposes; MoST and STyMo are trained per style.

Table 4: **Character grouping for the 3AFC block**, from ground-truth deltas without any model. Posture direction: cosine between a style's mean joint offset and the centroid of each group (the style itself left out). Limb dynamics: end-effector (hands and feet) speed and jerk of the styled take relative to the neutral take, averaged over choreographies. Each descriptor is split into two clusters; the two splits agree on every style except *Clumsy*, which is excluded.

	Posture direction			Limb dynamics		
Style	cos to A	cos to B	group	speed ratio	jerk ratio	group
Tough	+0.67	−0.61	A	1.48	4.22	A
Comic	+0.67	−0.25	A	1.14	3.21	A
Cute	−0.05	+0.04	B	0.76	2.11	B
Sexy	−0.62	+0.25	B	0.68	2.15	B
Chessy	−0.86	+0.34	B	0.49	1.67	B
Swag	−0.22	+0.44	B	0.57	1.79	B
Fresh	+0.22	+0.29	B	0.90	2.58	B
Clumsy	+0.13	−0.06	*A*	0.68	2.44	*B*

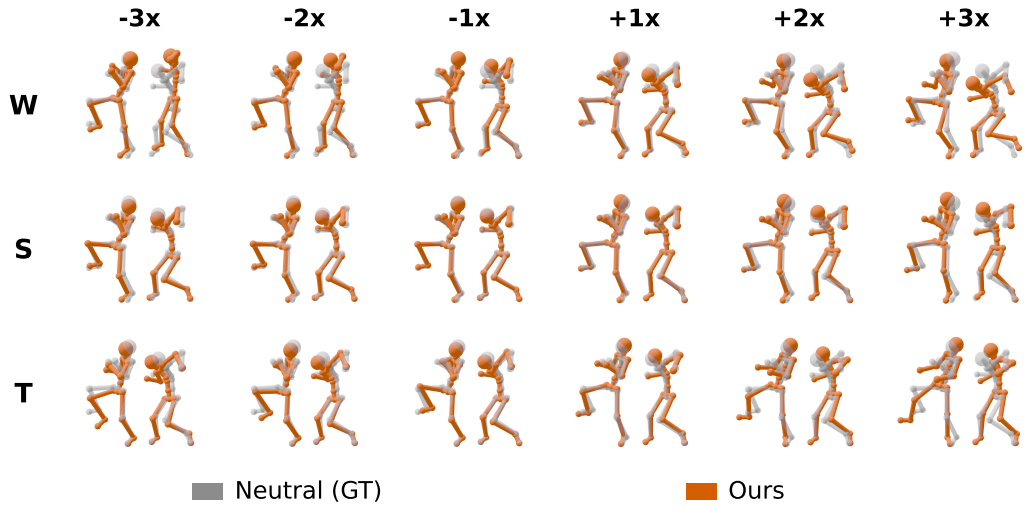

Figure 6: **Signed and graded control on the Effort axes.** Two poses of one choreography under Weight, Space and Time at six coordinate scales; grey is the neutral input and orange is the output, aligned at the root so the offset is the deformation. The sign reverses the direction and the scale grades it, with no retraining between cells: the coordinates are simply multiplied. Space moves the pose least, which is consistent with its measured weakness as an axis (Table 3).

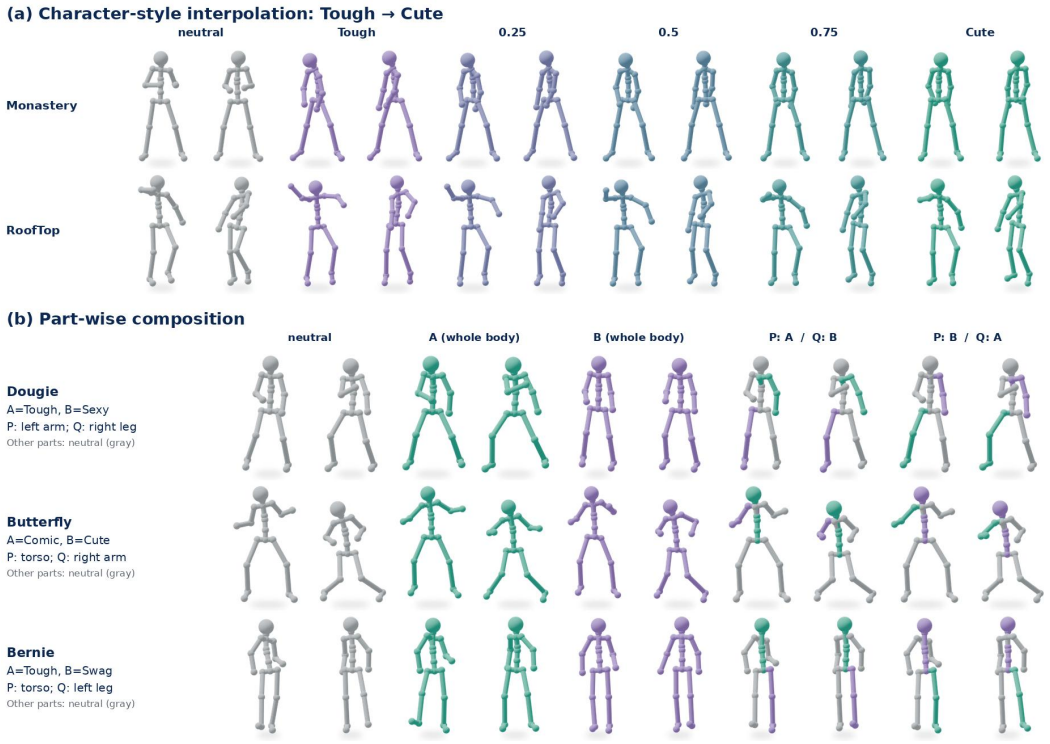

Figure 7: **Editing in coordinate space.** Top: interpolation between two character coordinates (Tough → Cute) on two unseen choreographies; the output moves continuously between the styles with no retraining. Bottom: part-wise composition on selected anatomical parts: left arm and right leg (Dougie), torso and right arm (Butterfly), and torso and left leg (Bernie). The last two columns swap styles A and B between parts P and Q; gray marks parts with no assigned style deformation before projection. Left and right refer to the performer. Two frames per cell.

F ABLATIONS

Tables 5–9 change one design choice at a time and report the quantity that choice affects (Sec. 4.6). Each table is preceded by what its rows mean and why the deployed setting was kept.

What the dance fields are (Table 5). The field bank f_θ is replaced while everything downstream, including the ridge solve, stays the same, so the table isolates what the fields contribute. A frozen random MLP has the same architecture and the same content input but no training, and the character cosine falls from 0.594 to 0.267: the ridge cannot compensate for fields that do not span the deformations, so the bank has to be learned. Static fields are learned but do not see the content, and they reach 0.386: a style is then one deformation for every choreography, which is exactly the assumption of Sec. 1 that dance breaks. Character transfer needs both, learned fields conditioned on the content; the Effort axes are less demanding (Weight is recoverable even from static fields), which is consistent with their fixed reading.

Table 5: **Ablation (a): what the dance fields are.** The field bank f_θ is replaced and the coordinates are solved as usual (character styles, contiguous 32-frame anchor, six folds). Character transfer needs learned, content-conditioned fields.

Fields	Content-conditioned	Learned	Character cosine
Learned MLP (ours)	✓	✓	**0.594**
Frozen random MLP	✓	–	0.267
Static fields	–	✓	0.386

How the coordinates are read (Table 6). With the fields fixed, the rows vary the three choices of the read-time solve. Centering the ridge on 0 instead of the constant $\bar{\mathbf{a}}_c$ is the decisive one: at 0, one keypose *replaces* the constant and falls to 0.333, below reference-free (0.450), because a single pose cannot determine 65 coordinates and the ridge shrinks the rest toward nothing; centered on the constant, the same pose *corrects* it and reaches 0.569, and the gain grows with every added pose. Solving per body part adds 0.01–0.02 at three and five poses, since a pose then constrains five small systems rather than one large one, and $K = 64$ over $K = 8$ adds 0.02–0.11, more at more poses, because a larger bank has more directions for the poses to select. The deployed row combines the three and is best at every keypose count; its 0-keypose value is 0.430 rather than 0.450 only because the part-wise constant is a different estimator, not because part-wise reading costs anything.

Table 6: **Ablation (b): how the coordinates are read.** Character styles, six folds, style direction cosine vs. number of keyposes. The 0-keypose column is one coordinate per style (the stored constant); every row that observes keyposes must beat it. Prior-centered ridge, part-wise solve and $K = 64$ each contribute, and the deployed setting (last row) is best at every anchor count.

Solve	Center	Part-wise	K	0	1	3	5
Ridge	0	–	64	0.450	0.333	0.574	0.668
Ridge	prior	–	8	0.441	0.549	0.600	0.617
Ridge	prior	–	64	0.450	0.569	0.667	0.708
Ridge	0	✓	64	0.430	0.482	0.567	0.601
Ridge (ours)	prior	✓	64	0.430	**0.578**	**0.683**	**0.730**

The remaining choices (Table 7). *Deformation space.* The same model trained on rotation residuals in the tangent space of joint rotations reaches a character cosine of 0.150 (oracle 0.301) against 0.711 (oracle 0.918) in joint-position space: a style that moves the hands is a small rotation at the shoulder and a large one at the wrist, so the target is spread across the chain in a way the fields do not capture, whereas in position space it is one displacement at one joint. *Style constant.* Averaging per-choreography coordinates cancels magnitude—the Weight deformation shrinks from 3.68 to 1.18 cm—so the magnitude correction restores it to the ground-truth size; the averaged constant also beats a learned per-condition head (0.430 vs. 0.413), and a constant tuned directly on the evaluation metric lands at cosine 0.986 to the average while doing worse on test, so there is no better constant

to find. *Keypose alignment.* A 5-frame offset (3.6% of the clip) drops the Effort cosine from 0.607 to 0.222, below reference-free, and re-attaching the poses restores 0.596: what a misplaced pose loses is its correspondence, not its information, which is the alignment requirement of Appendix A as a measurement. *Prior sample size.* The constant averaged over 14 or more training choreographies is 0.448 / 0.406 (character / Weight); over 3 it is 0.322 / 0.243, over 1 it is 0.251 / 0.164, so the reference-free path needs the breadth of the training set, not one good example. *Keypose noise.* Gaussian noise of 1–2 cm on every joint of the keyposes changes the cosine by at most 0.014, and only ≈ 20 cm pushes the reading below the constant: the solve is far more tolerant of where a joint is than of when it is. *Time axis.* With local tempo removed ($\eta \equiv 0$) the two poles of Time differ by 0.10 cm in pose against 7.82 cm with it, so global tempo alone changes the length but not where the performance rushes and lingers, which is what Sudden and Sustained are. *Temporal smoothing.* A Gaussian of $\sigma = 2$ frames on $\mathbf{\Delta}$ brings ΔJerk from $+120\%$ to $+14\%$ at no cost in cosine, and is the only post-processing applied.

Table 7: **Ablation (c): representation, style constant, alignment, prior, noise, time axis and smoothing.** Within each block the rows are alternatives for one design choice; ✓ marks the deployed one, and the block reports the quantity that choice affects. Keypose noise pushes the reading below the reference-free constant only at ≈20 cm.

Design choice	Variant	Ours	Measured
Deformation space	position space	✓	character cosine **0.711** (oracle 0.918)
	rotation tangent space	–	character cosine 0.150 (oracle 0.301)
Style constant $\bar{\mathbf{a}}_c$	averaged, magnitude-corrected	✓	Weight deformation **3.68** cm (GT 3.68)
	averaged only	–	Weight deformation 1.18 cm
	learned per-condition head	–	reference-free cosine 0.413 (averaged: **0.430**)
	optimized on the evaluation metric	–	cosine 0.986 to the average; lower on test
Keypose alignment	keyposes at their true moment	✓	Effort cosine **0.607**
	offset by 5 frames (3.6% of clip)	–	Effort cosine 0.222
	re-attached after the offset	–	Effort cosine 0.596
Prior sample size	averaged over ≥ 14 training choreographies	✓	char. / Weight cosine **0.448** / **0.406**
	3 choreographies	–	char. / Weight cosine 0.322 / 0.243
	1 choreography	–	char. / Weight cosine 0.251 / 0.164
Keypose joint noise	none	✓	—
	σ = 1–2 cm on every joint	–	cosine change -0.001 to -0.014
Time axis	local + global tempo	✓	pose difference between T poles **7.82** cm
	global tempo only ($\eta \equiv 0$)	–	pose difference between T poles 0.10 cm
Temporal smoothing of $\mathbf{\Delta}$	Gaussian $\sigma = 2$	✓	ΔJerk **+14** %
	none	–	ΔJerk +120 %

Whole-body training, part-wise reading. The coordinates are solved whole-body during training ($\lambda = 0.01$, dozens of support frames) but part-wise at inference ($\lambda_p = 30$, k poses). Table 8 asks whether that mismatch costs anything, on fold 0 (four held-out choreographies). Part-wise reading is what matters: with the deployed fields, five keyposes read part-wise reach 0.536 / 0.509 (character / Effort) against 0.510 / 0.451 read whole-body, and the gap widens with the number of keyposes (none at one pose), which is the six-fold effect of Table 6. Training part-wise as well adds only $+0.013$ / $+0.011$, within seed variance, and makes the fields depend on the reading: the same part-wise-trained model read whole-body falls to 0.462 / 0.398. Whole-body training is therefore kept; it is robust to either reading, and the part-wise gain is obtained at read time.

Table 8: **Ablation (d): where the coordinates are solved part-wise.** Style direction cosine vs. number of keyposes, character / Effort, fold 0 (four held-out choreographies), prior-centered ridge $\lambda_p = 30$ in both readings. Rows pair the solve used in training with the solve used to read keyposes at inference.

Training solve	Reading solve	1 keypose	3 keyposes	5 keyposes
whole-body (ours)	**part-wise (ours)**	0.421 / 0.348	0.508 / 0.462	0.536 / 0.509
whole-body	whole-body	0.421 / 0.343	0.510 / 0.442	0.510 / 0.451
part-wise	part-wise	0.439 / 0.362	0.518 / 0.480	0.549 / 0.520
part-wise	whole-body	0.388 / 0.301	0.455 / 0.385	0.462 / 0.398

The read-time ridge strength. λ_p decides how far k poses may move the coordinates away from the constant $\bar{\mathbf{a}}_c$, and Table 9 sweeps it over three orders of magnitude. Both failure modes are visible. Too small, and the solve fits the k poses alone: at $\lambda_p = 1$ the character cosine drops by 0.02 with one pose and the Effort cosine by 0.04 with three, the poses over-determining nothing but their own frames. Too large, and the poses stop mattering: at $\lambda_p = 1000$ the cosine is already 0.12 below the optimum and falling toward the reference-free value. Between them the optimum is flat from 10 to 30, and it is the same for character and Effort styles and for one and three keyposes, so λ_p needs no tuning per style or per keypose count. We use 30, the upper end of the plateau, which leans on the prior rather than on the poses when the two disagree.

Table 9: **Ablation (e): the read-time ridge strength λ_p.** Style direction cosine, six folds, part-wise prior-centered ridge (Sec. 3.2), character / Effort. $\lambda_p \to \infty$ recovers reference-free (0.451 / 0.157). The optimum is flat over 10–30; we use 30.

λ_p	1	3	10	**30**	100	300	1000
1 keypose	0.565 / 0.462	0.577 / 0.473	0.585 / 0.482	**0.585 / 0.483**	0.571 / 0.470	0.544 / 0.448	0.495 / 0.408
3 keyposes	0.667 / 0.543	0.681 / 0.565	0.689 / 0.578	**0.689 / 0.580**	0.672 / 0.564	0.636 / 0.529	0.572 / 0.471

G TABLE 1 WITH CONFIDENCE INTERVALS

Table 10: Choreography-level means of Table 1 with choreography-clustered bootstrap 95% confidence intervals ($n = 29$). CC/SC in cm, PFC $\times 10^{-3}$.

	Effort			Character		
Method	CC↓	SC↓	PFC↓	CC↓	SC↓	PFC↓
GANimator	83.4 [79.9, 86.9]	85.8 [82.7, 88.8]	0.91 [0.75, 1.06]	80.4 [77.2, 83.6]	83.7 [81.3, 86.2]	1.50 [1.17, 1.87]
SinMDM	106.0 [101.8, 110.6]	105.5 [101.7, 109.6]	1.26 [1.12, 1.40]	96.6 [92.7, 100.9]	95.5 [92.5, 98.8]	1.80 [1.45, 2.20]
MoST	70.6 [67.3, 73.6]	78.7 [75.5, 81.8]	0.93 [0.79, 1.07]	70.8 [68.1, 73.4]	79.4 [76.8, 81.9]	1.57 [1.19, 2.01]
MoMo	87.3 [82.9, 91.5]	91.2 [86.8, 95.3]	3.30 [2.66, 3.96]	88.6 [84.7, 92.3]	89.8 [86.4, 93.2]	5.04 [4.04, 6.13]
STyMo	54.4 [52.4, 56.3]	63.0 [61.0, 65.1]	0.34 [0.22, 0.48]	60.4 [57.7, 63.2]	71.1 [68.5, 73.8]	0.41 [0.28, 0.54]
Ours, 0 examples	42.8 [41.8, 43.9]	52.3 [50.2, 54.4]	0.30 [0.24, 0.37]	66.2 [63.8, 68.7]	69.2 [66.2, 72.1]	0.99 [0.73, 1.27]
Ours, 1 keypose	30.7 [28.8, 32.6]	43.1 [40.5, 45.5]	0.26 [0.20, 0.34]	46.2 [42.6, 50.1]	53.9 [50.6, 57.8]	0.67 [0.46, 0.89]
Ours, 3 keyposes	32.2 [30.2, 34.2]	40.2 [37.7, 42.6]	0.25 [0.20, 0.32]	48.2 [45.2, 51.5]	48.6 [45.6, 51.7]	0.67 [0.47, 0.88]
Ours, 5 keyposes	31.0 [28.9, 33.1]	37.2 [34.7, 39.7]	0.26 [0.20, 0.35]	47.9 [44.7, 51.0]	44.8 [41.7, 48.0]	0.71 [0.49, 0.94]

H TRANSFER BEYOND THE TRAINING DATA

The fields are trained on our dance data only. Figure 8 applies them unchanged to two sources that never appear in training, and Table 11 measures the first of them.

MOCHA: unseen rig, actions and styles. The 25-joint MOCHA rig, its locomotion actions and its eight styles are all outside the training data. Each style is read from the one paired clip MOCHA provides per style, either from keyposes of it or from the whole clip, without any training, whereas STyMo, GANimator and SinMDM train a model on that pair. In the top rows of Fig. 8 the style is nevertheless applied: the styled poses differ from the neutral input in the direction of the reference pair, and the content action is kept. Table 11 confirms this on the full protocol. Ours preserves content best (CC 49.7 cm from the whole clip, 53.3 from three keyposes, against STyMo's 62.8), has the best physical foot-contact score (PFC 0.029 against 0.051), and keeps jerk in STyMo's range. SOC is the one column where STyMo is ahead, and the reason is what each method knows about the reference clip: STyMo fits its static posture model to that clip and so reproduces its mean posture offset almost exactly (cosine 0.82, 95% of its magnitude), whereas ours never trains on the clip—it reads a few of its poses through a ridge solve centered on a dance prior—and reproduces the offset partially (cosine 0.47–0.53, 69–90% of the magnitude from three keyposes to the full pair), still well ahead of every other baseline, whose offsets point elsewhere (MoST -0.09).

AIST++: familiar movements, unseen performances. Panel (b) of Fig. 8 applies our own styles to AIST++ performances of choreographies that also appear in our data, danced by other dancers

in another capture. Character styles are read from three keyposes and the Effort axes are applied reference-free; in both cases the style appears on the new performance, which shows that the fields do not depend on the dancer or the capture they were trained on.

AIST++: unseen performances and unseen genres. The bottom rows apply our own styles to AIST++ performances. Some are choreographies that also appear in our data, danced by other dancers in another capture: character styles are read from three keyposes and the Effort axes are applied reference-free, and in both cases the style appears on the new performance, which shows that the fields do not depend on the dancer or the capture they were trained on. Others are from genres absent from training—ballet, house, breaking—on which the Effort axes still transfer and remain controllable by scaling the coordinate.

Table 11: MOCHA protocol (Ponton et al., 2026): 17 neutral AverageJoe clips $\times$ 8 styles, every method re-measured in one canonical joint space. CC is MoST's mean per-frame L2 to the content (cm), PFC EDGE's physical foot-contact score ($\times 10^{-3}$), ΔJerk the change in jerk relative to the source (%), and SOC the cosine between the output's mean posture offset and the reference pair's. Baselines use the source root, STyMo its own. Ours rows read the coordinates from keyposes of the one paired clip per style, or from the whole clip (the data that per-style models are trained on), still without training. Best in bold, second underlined.

	Style data	CC $\downarrow$	PFC $\downarrow$	ΔJerk $\rightarrow 0$	SOC $\uparrow$
Identity floor	—	0.0	0.017	0.0	—
GANimator (Li et al., 2022)	per-style	66.1	0.084	412.7	0.296
SinMDM (Raab et al., 2024b)	per-style	97.4	0.185	258.6	0.455
MoST (Kim et al., 2024)	released	79.2	0.084	493.9	-0.088
MoMo (Raab et al., 2024a)	released	149.2	0.368	1312.5	0.110
AStF	released	83.1	0.167	967.7	0.101
STyMo (Ponton et al., 2026)	per-style, 1 pair	62.8	0.051	170.8	**0.822**
Ours	3 keyposes	<u>53.3</u>	0.035	204.2	0.473
Ours	5 keyposes	55.1	<u>0.031</u>	190.6	0.501
Ours	full pair, $\gamma = 0.7$	**49.7**	**0.029**	<u>163.7</u>	0.506
Ours	full pair, $\gamma = 1.0$	70.7	0.034	**149.5**	<u>0.528</u>

REFERENCES

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics*, 39(4), 2020.

Janet Adshead-Lansdale and June Layson. *Dance history: An introduction*. Routledge, 2006.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations (ICLR)*, 2023.

Adina Armelagos and Mary Sirridge. The identity crisis in dance. *The Journal of Aesthetics and Art Criticism*, 37(2):129–139, 1978.

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations (ICLR)*, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Susan Leigh Foster. *Reading dancing: Bodies and subjects in contemporary American dance*. Univ of California Press, 1986.

Nelson Goodman. *Languages of art: An approach to a theory of symbols*. Hackett publishing, 1976.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 9318–9333, 2023.

Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, 35(4), 2016.

Lei Hu, Zihao Zhang, Yongjing Ye, Yiwen Xu, and Shihong Xia. Diffusion-based human motion style transfer with semantic guidance. In *Computer Graphics Forum*, volume 43, pp. e15169. Wiley Online Library, 2024.

Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics*, 41(3), 2022.

Deok-Kyeong Jang, Yuting Yang, Andreas Aristidou, and Sung-Hee Lee. Mocha: Real-time motion characterization via context matching. *ACM SIGGRAPH Asia*, 2023.

Bin Ji, Ye Pan, Zhimeng Liu, Shuai Tan, Xiaogang Jin, and Xiaokang Yang. Pomp: Physics-constrainable motion generative model through phase manifolds. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22690–22701. IEEE, 2025.

Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. Most: Motion style transformer between diverse action contents. In *CVPR*, 2024.

Rudolf Laban and Lisa Ullmann. The mastery of movement. 1971.

Roderyk Lange. Laban and movement notation. In *The Laban Sourcebook*, pp. 155–166. Routledge, 2012.

Susanne K Langer and . Langer. *Feeling and form*, volume 3. Routledge and Kegan Paul London, 1953.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. Ganimator: Neural motion synthesis from a single sequence. *ACM Transactions on Graphics*, 41(4), 2022.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13381–13392. Ieee, 2021.

Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proc. ACM Comput. Graph. Interact. Tech.*, 5(1), 2022.

Soomin Park, Deok-Kyeong Jang, and Sung-Hee Lee. Diverse motion stylization for multiple style domains via spatial-temporal graph-based generative model. *Proc. ACM Comput. Graph. Interact. Tech.*, 4(3), 2021.

Jose Luis Ponton, Alexander Winkler, Ladislav Kavan, Yuting Ye, and Petr Kadlecek. Stymo: Fast and controllable few-shot motion style transfer. *ACM Transactions on Graphics (TOG)*, 45(4): 1–13, 2026.

Sigal Raab, Inbar Gat, Nathan Sala, Guy Tevet, Rotem Shalev-Arkushin, Ohad Fried, Amit H. Bermano, and Daniel Cohen-Or. Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer. In *ACM SIGGRAPH Asia*, 2024a.

Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H. Bermano, and Daniel Cohen-Or. Single motion diffusion. In *ICLR*, 2024b.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *International Conference on Learning Representations (ICLR)*, 2020.

Maxine Sheets-Johnstone. *The phenomenology of dance*. Temple University Press, 2015.

Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 821–830. IEEE, 2024.

Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (ToG)*, 41(4):1–13, 2022.

Xiangjun Tang, Linjun Wu, He Wang, Yiqian Wu, Bo Hu, Songnan Li, Xu Gong, Yuchen Liao, Qilong Kou, and Xiaogang Jin. Decoupling contact for fine-grained motion style transfer. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Tianxin Tao, Xiaohang Zhan, Zhongquan Chen, and Michiel van de Panne. Style-erd: Responsive and coherent online motion style transfer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6583–6593. IEEE, 2022.

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 448–458, 2023.

Linjun Wu, Xiangjun Tang, Jingyuan Cong, He Wang, Bo Hu, Xu Gong, Songnan Li, Yuchen Liao, Yiqian Wu, Chen Liu, et al. Semantically consistent text-to-motion with unsupervised styles. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025.

Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics*, 34(4), 2015.

Wenjie Yin, Hang Yin, Kim Baraka, Danica Kragic, and Mårten Björkman. Dance style transfer with cross-modal transformer. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5058–5067, 2023.

Lei Zhong, Yiming Xie, Varun Jampani, Deqing Wei, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *ECCV*, 2024.

Luisa Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning (ICML)*, 2019.



Figure 8: **Transfer beyond the training data.** **(a)** MOCHA: unseen rig, actions, and styles, with three-keypose style references. **(b)** AIST++: unseen performances of movements present in our dataset—RunningMan, ShamRock, and RogerRabbit—with character transfer from three keyposes and reference-free W/S controls at $+2$. **(c)** AIST++: movements absent from our training dataset, with reference-free W/S/T controls at ± 2. Two frames per cell; (c) uses the same elapsed times (1 and 2 seconds) and display scale across each row, so T shows timing changes. The fields remain fixed in all panels.