LEARNING STYLE, REUSING MOTION: PRIOR-GUIDED UNPAIRED MOTION STYLIZATION

Anonymous authors
Paper under double-blind review

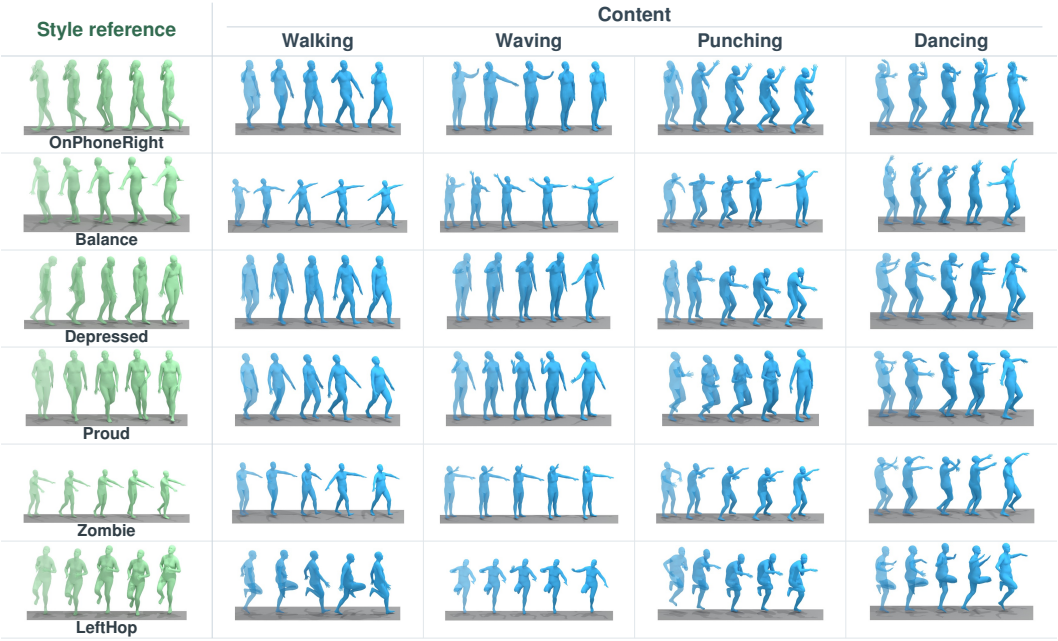

Figure 1: TriPrior combines four content texts with six style motions. Green meshes show reference motions; blue meshes show stylized motions. Sampled poses progress from light to dark over time and are spread horizontally for visibility; each case is uniformly scaled. More visualization results are available at `https://triprior.github.io`.

ABSTRACT

Motion stylization controls how a text-specified motion is performed while preserving its content. Paired performances of the same content in different styles are scarce, limiting direct supervision. Pretrained text-to-motion generators offer broad content capabilities and motion knowledge, but style edits may conflict with both. Our framework, **TriPrior**, coordinates two data priors from unpaired motions with the frozen generator's own motion prior. These priors act at inference time by correcting flow-model velocity predictions or, equivalently, diffusion-model score predictions. The dataset-level prior supplies a stable style direction, the instance-level prior adapts it to the generated motion, and the generative prior provides content-conditioned correction through the generator's response to the edited sampling state. Keeping the data priors outside the generator enables their reuse on HY-Motion, HY-Lite, and Kimodo without backbone retraining. On HumanML3D with style-motion input, TriPrior on HY-Motion achieves 29.40% average accuracy across four independent style classifiers (SRA-4) and 38.70% content-retrieval R@1, exceeding MoMo and MLD+Aberman by 3.06 and 11.32 percentage points, respectively. Controlled ablations show that the generative prior improves content retrieval and motion quality while largely preserving style recognition across the evaluated styles and content motions.

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

3D human motion brings characters to life in games, films, and virtual worlds. Text-to-motion generation creates it efficiently from natural language (Guo et al., 2022; Chen et al., 2023). Yet compelling characters depend on both what they do and how: the same walk can appear tired, proud, or zombie-like, creating different character impressions. Motion stylization is therefore essential for vivid character animation: it preserves text-specified motion content while adapting performance to the intended character (Figure 1). Here text specifies content, and a reference motion or style name specifies the target style (Zhong et al., 2024; Kim et al., 2025; Guo et al., 2025b). While learning this transformation via supervised training requires paired motions that express the same content in different styles, collecting such pairs at scale is infeasible in practice.

Recent methods already support motion stylization without requiring paired performances of the same content across styles. They commonly introduce style control through backbone-specific modules, inference-time guidance, or a combination of both (Raab et al., 2024; Zhong et al., 2024; Guo et al., 2025b; Kim et al., 2025). Two challenges remain. First, modules tied to a particular architecture, motion representation, or sampling process require redesign and retraining for each new backbone. This adaptation becomes increasingly costly for large generators such as HY-Motion and Kimodo (Wen et al., 2025; Rempe et al., 2026). Second, inference-time editing requires deciding not only how strongly to express a style, but also whether a proposed change is compatible with the specified content and the generator's motion prior. Existing style objectives can encourage recognizable style, yet provide limited evidence of this compatibility. Weak edits may therefore underexpress the style, whereas strong or misdirected edits can introduce reference content, degrade text alignment, or produce implausible motion. We seek reusable style control that retains the flexibility of inference-time editing, while the pretrained generator identifies and corrects unsuitable edits before they compromise content or motion quality.

To address these challenges, in this paper we propose a novel and effective inference-time editing framework, referred to as **TriPrior**, for motion stylization without paired motions. Since representative pre-trained text-to-motion models are mainly flow-based models and diffusion models, TriPrior achieves motion stylization by adjusting velocities in the velocity fields of flow-based models, and equivalently, scores in the score space of diffusion models. This design makes TriPrior a backbone-agnostic framework compatible with different pretrained text-to-motion models as shown in the experiments. The core component behind TriPrior is a series of style priors mined from both unpaired data and the underlying pre-trained model, which synergically work together while being complementary to each other. Specifically, TriPrior at first computes a *dataset-level data prior* from unpaired data, which measures the averaged difference between a set of stylized motions and a set of neutral motions as a stable overall trend. This prior is used during the beginning stage of a generation process, adjusting estimated velocities towards a stylized destination. At a later stage of the generation process, TriPrior turns to an *instance-level data prior* and a *generative prior* for fine-grained and content-preserving adjustment. The instance-level data prior is implemented as a style classifier that predicts the style of a specific motion, so that its feedback can be used as fine-grained style control signals. Although effective, using this strong prior alone may degrade content quality because it emphasizes style alignment. Thus the generative prior is further adopted in TriPrior to retain content quality by reusing velocity estimates from the vanilla pre-trained text-to-motion model. The key observation is that when the current sampling state may lead to a degraded output, the pre-trained model itself will *estimate a bounded correction that brings the state back* due to the intrinsic properties of flow-based models. These two priors are applied in an interleaved manner to balance between stylization and content preservation.

Together, these three priors establish a coarse-to-fine division of labor. The dataset-level prior provides a stable style direction, the instance-level prior adapts it to the motion being generated, and the generative prior regulates whether the resulting edits remain compatible with the specified content and plausible motion dynamics. This formulation separates learning how a motion should be performed from relearning what motion should occur: unpaired style data provide the former, while the pretrained generator retains the latter from large-scale training. Moreover, because the data priors remain external to the generator and the generative prior is accessed through its velocity or score predictions, the same style controller can be connected to different backbones without retraining their internal architectures. TriPrior therefore allows motion stylization to reuse both existing style knowledge and the increasingly capable priors of pretrained motion generators.

Experiments on HumanML3D validate both the complementary roles of the three priors and their reuse across backbones. With style-motion input, TriPrior on HY-Motion achieves 29.40% SRA-4—the mean accuracy of four heterogeneous style classifiers independent of generation guidance—and 38.70% content-retrieval R@1. Compared with the strongest relevant baselines, these results improve style recognition by 3.06 percentage points over MoMo and content retrieval by 11.32 points over MLD+Aberman, showing that the dataset- and instance-level priors strengthen style expression while retaining the requested content. The same data priors transfer without retraining to HY-Lite and Kimodo. Controlled HY-Motion experiments further show that the generative prior improves distributional and physical quality after style editing while largely preserving style recognition and content retrieval across the evaluated styles.

Our main contributions are:

1. We introduce TriPrior, a backbone-agnostic inference-time editing framework that stylizes flow- and diffusion-based motion generators through corrections to their velocity or score predictions, without paired motions or backbone retraining.
2. We formulate stylization through three complementary priors: dataset- and instance-level priors extract stable and motion-specific style control from unpaired data, while the frozen generator provides a generative prior that corrects edits for content and motion compatibility.
3. Experiments on HumanML3D demonstrate stronger joint style and content performance than prior methods, reuse of the same data priors across HY-Motion, HY-Lite, and Kimodo, and the corrective effect of the generative prior.

2 Related Work

Early data-driven motion stylization learned motion manifolds or feed-forward style mappings from structured motion collections (Holden et al., 2016; 2017). Subsequent work reduced supervision through few-shot adaptation and online transfer (Mason et al., 2018; Tao et al., 2022), and separated content from style using unpaired motions (Aberman et al., 2020; Jang et al., 2022); MoST further transferred style across different action contents (Kim et al., 2024). Other approaches unify synthesis and style transfer with diffusion, disentangle style in optimized or quantized latent representations, or correct cross-dataset transfer in a second stage (Chang et al., 2022; Gupta et al., 2024; Guo et al., 2025a; Zargarbashi et al., 2026). With the emergence of text–motion datasets and diffusion generators (Guo et al., 2022; Tevet et al., 2023; Chen et al., 2023), stylization has increasingly been incorporated into text-conditioned generation. SMooDi adapts frozen MLD through an additional ControlNet-style branch and classifier-feature guidance (Zhong et al., 2024). This design provides effective style conditioning, but its learned branch is tied to the representation and architecture of MLD; its measured style accuracy can also depend strongly on the evaluator when guidance and evaluation favor related classifier features. MoMo instead reuses diffusion attention for training-free transfer (Raab et al., 2024). Recent methods broaden style conditioning through bidirectional control, cross-modal fusion, multiple references, and body-part text reasoning (Li et al., 2024; Guo et al., 2025b; Kim et al., 2025; Chen et al., 2026; Zhong et al., 2025). Together, these advances reduce paired supervision and expand the available style inputs, while their control mechanisms generally remain coupled to a particular motion representation, generator architecture, or sampling process.

Text-driven motion generation has progressed from MDM and MLD to larger models such as HY-Motion and Kimodo, which use different representations and samplers (Tevet et al., 2023; Chen et al., 2023; Wen et al., 2025; Rempe et al., 2026). Stylization methods built inside a fixed generator require renewed adaptation when the backbone changes, with training cost potentially increasing with model scale (Zhong et al., 2024; Li et al., 2024; Guo et al., 2025b; Kim et al., 2025). Other methods manipulate samples at inference time using classifier features or pretrained attention (Zhong et al., 2024; Raab et al., 2024). Such control can strengthen style, but its style signal alone provides limited evidence that an edit remains compatible with the requested content and the generator's motion knowledge. TriPrior addresses these two issues through complementary external and generative priors. Dataset- and instance-level priors extracted from unpaired motions remain outside the backbone and specify how the generated motion should express a style. The frozen generator itself supplies the generative prior that regulates these edits using its existing content and motion knowledge. Accessing these priors through velocity or score predictions allows the same style controller to operate on flow- and diffusion-based generators without retraining their internal architectures.

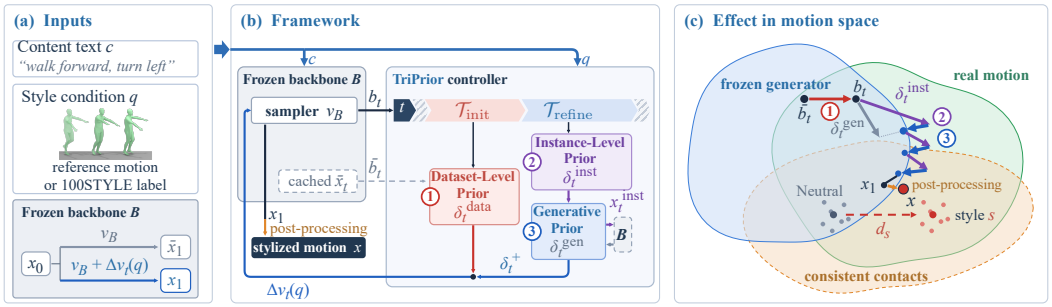

Figure 2: Overview of TriPrior. **(a)** Content text c and style condition q condition two trajectories from the same noise: the frozen backbone gives the cached content trajectory, while TriPrior adds $\Delta v_t(q)$ to produce the stylized trajectory. **(b)** The backbone predicts clean motion b_t, and the TriPrior controller coordinates three components. For $t \in \mathcal{T}_{\text{init}}$, the **Dataset-Level Prior** applies a relative style direction. For $t \in \mathcal{T}_{\text{refine}}$, the **Instance-Level Prior** proposes δ_t^{inst}, and the **Generative Prior** obtains the response δ_t^{gen} through an extra backbone query; their combination gives δ_t^{+}. The resulting clean-motion edit is converted to $\Delta v_t(q)$, and hatched cells denote inactive sampling steps. Grey denotes the frozen backbone, and dashed boxes denote its cached trajectory and extra query. **(c)** In motion space, the Dataset-Level Prior initializes style relative to the cached prediction, the Generative Prior modulates later Instance-Level proposals under the original content condition, and lightweight post-processing runs once after sampling.

3 Method

Learning stylization inside a pretrained generator would tie style control to a particular backbone, while style-driven edits alone cannot determine whether they remain compatible with the requested content and plausible motion dynamics. TriPrior therefore freezes the backbone and coordinates three complementary priors during sampling (Figure 2). In the early stage, the dataset-level prior initializes a stable style direction from unpaired motions. In the later stage, the instance-level prior proposes an edit adapted to the current motion, and the generative prior immediately corrects it using the frozen backbone. This proposal–correction pair repeats across the active late-stage sampling nodes, allowing style expression and pretrained motion regularization to interact throughout refinement. A lightweight post-processing stage removes residual artifacts after sampling. Sections 3.2–3.4 describe the three priors, while Section 3.5 connects their shared edit interface to different backbones. Algorithm 1 summarizes the complete inference procedure.

3.1 Task and Notation

Given content text c and a style condition q, provided as either a reference motion or a closed-set 100STYLE name (Mason et al., 2022), TriPrior stylizes the sampling trajectory of a frozen text-to-motion backbone B without updating its parameters.

Taking a flow-matching backbone as an example, we compare its original content-conditioned trajectory with the stylized trajectory produced by TriPrior. Let $\mathcal{T}_{\text{init}}$ and $\mathcal{T}_{\text{refine}}$ denote the initialization and refinement portions of the sampling trajectory, respectively:

$$\begin{aligned} \text{Pretrained T2M:} \quad & \dot{\bar{x}}_t = v_B(\bar{x}_t, t, c), \qquad \bar{x}_0 \sim \mathcal{N}(0, I), \\ \text{TriPrior:} \quad & \dot{x}_t = v_B(x_t, t, c) + \Delta v_t(q), \qquad x_0 = \bar{x}_0, \\ \Delta v_t(q) = & \begin{cases} \Delta v_t^{\text{data}}, & t \in \mathcal{T}_{\text{init}}, \\ (1-\beta)\Delta v_t^{\text{inst}} + \beta \Delta v_t^{\text{gen}}, & t \in \mathcal{T}_{\text{refine}}, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{1}$$

Here $\beta \in [0, 1]$ blends the instance-level proposal and generator response during refinement. Both trajectories start from the same noise. The cached content trajectory $\bar{x}_t$ follows only the content-conditioned velocity field v_B, whereas the stylized trajectory x_t additionally receives $\Delta v_t(q)$ determined by style condition q. Their final states are the cached content motion $\bar{x}_1$ and the stylized motion $x = x_1$, respectively. At an intermediate time, x_t is the current sampling state and may still

contain noise; b_t is its clean-motion prediction, and $\bar{b}_t$ is the corresponding prediction along the cached trajectory. The three priors evaluate and edit b_t, then convert the edit back into a velocity correction that changes the trajectory of x_t. For any prior k, the clean-motion edit and its velocity correction satisfy the following relationship:

$$\begin{aligned} b_t &= x_t + (1-t)v_B(x_t, t, c), \\ b_t + \delta_t^k &= b_t + (1-t)\Delta v_t^k \quad \Longrightarrow \quad \Delta v_t^k = \frac{\delta_t^k}{1-t}, \end{aligned} \tag{2}$$

where $k \in \{\mathrm{data}, \mathrm{inst}, \mathrm{gen}\}$ indexes the dataset-, instance-, and generative-prior edits, respectively. Thus, changing the predicted clean motion by δ_t^k requires the velocity correction used in Equation 1. Sections 3.2–3.4 define these three edits in turn.

3.2 Dataset-Level Prior: Relative Style Direction

The dataset-level prior first isolates a content-balanced style direction from unpaired motions, then applies it relative to the cached content trajectory. The first step separates style from content-dependent motion statistics; the second prevents an absolute style target from overwriting generated content.

Before sampling, a style name directly specifies the target class s_q, whereas a reference motion is assigned to s_q by a frozen reference-style recognizer. This input mapping is separate from the classifier guidance in Section 3.3. Let $\phi(\cdot)$ be a fixed feature map that extracts style-relevant motion statistics, and let $\mu_{s,m}$ be the mean of $\phi(\cdot)$ over real motions with style s and content category m. Absolute style targets entangle style with content, so we subtract the Neutral mean within each category and average the differences to obtain the dataset-level direction:

$$d_s = \frac{1}{|\mathcal{M}_s|} \sum_{m \in \mathcal{M}_s} \left(\mu_{s,m} - \mu_{\mathrm{N},m}\right), \tag{3}$$

where N denotes Neutral and $\mathcal{M}_s$ contains the available content categories. Category-wise subtraction suppresses content bias, while averaging retains a stable style trend.

Because d_s is an average change rather than a content-specific motion, we match it against the cached content trajectory generated from the same text and noise. Using the same feature map ϕ, the initialization edit is defined by the following relative objective:

$$\delta_t^{\mathrm{data}} = \arg\min_{\delta} \left\| \frac{\phi(b_t + \delta) - \phi(\bar{b}_t) - \kappa d_{s_q}}{\sigma} \right\|^2, \qquad \Delta v_t^{\mathrm{data}} = \frac{\delta_t^{\mathrm{data}}}{1-t}. \tag{4}$$

Here δ is the optimization variable: a candidate motion edit in the shared motion representation. During each update, b_t, $\bar{b}_t$, ϕ, and d_{s_q} remain fixed. The scalar κ controls style strength, and σ normalizes feature scales. Bounded gradient steps approximate the minimizer δ_t^{data}, which is mapped to $\Delta v_t^{\mathrm{data}}$ for sampling. The relative objective expresses style as an edit to the generated content motion rather than as an absolute motion target.

3.3 Instance-Level Prior: Classifier Guidance

The dataset-level prior applies the same averaged direction to generated motions with different content. When $t \in \mathcal{T}_{\mathrm{refine}}$, the instance-level prior instead evaluates the current clean-motion prediction b_t with a frozen classifier trained on real motions. Because its gradient is computed locally at b_t, the same target style can induce different edits for walking, jumping, or kicking. Activating it after initialization lets a stable overall style precede motion-adaptive refinement.

For target class s_q, the classifier obtains the instance-level edit through

$$\begin{aligned} \mathcal{L}_{\mathrm{sty}}(\delta) &= -\log p_\psi(s_q \mid b_t + \delta), \\ \delta_t^{\mathrm{inst}} &= \arg\min_{\delta} \mathcal{L}_{\mathrm{sty}}(\delta), \qquad \Delta v_t^{\mathrm{inst}} = \frac{\delta_t^{\mathrm{inst}}}{1-t}. \end{aligned} \tag{5}$$

The gradient is again taken with respect to δ at the current b_t, making δ_t^{inst} motion specific. The classifier measures only target-style evidence, so the resulting proposal may still conflict with the requested content or pretrained motion prior. Section 3.4 corrects it before converting the accepted edit into the velocity update for that sampling step.

3.4 Generative Prior: Modulating Style Edits

The instance-level prior locally optimizes target-style evidence, but it does not account for how the pretrained generator propagates the resulting proposal under the original content condition. We therefore query the frozen backbone again from the edited sampling state at the same time t and content text c. The prediction change yields a content-conditioned response that is combined with the proposal, letting the frozen generator modulate the final update.

Let x_t^{inst} denote the sampling state obtained by applying δ_t^{inst} through Equation 2, and let $B(\cdot; t, c)$ denote the frozen backbone prediction in the shared clean-motion representation. The response and resulting edit are defined as follows:

$$\begin{aligned} \delta_t^{\text{gen}} &= B(x_t^{\text{inst}}; t, c) - B(x_t; t, c), \\ \delta_t^{+} &= (1-\beta)\delta_t^{\text{inst}} + \beta\delta_t^{\text{gen}}, \qquad \Delta v_t(q) = \frac{\delta_t^{+}}{1-t}. \end{aligned} \tag{6}$$

The two backbone queries share t and c and differ only by the instance-level proposal, so their difference defines δ_t^{gen}. Blending it with δ_t^{inst} gives the final edit δ_t^{+}, which is converted to $\Delta v_t(q)$ and applied to sampling. Because the response comes from the frozen generator under the original content condition, it lets the pretrained dynamics modulate the proposal's direction and magnitude before sampling; β controls the modulation strength. Each nonzero proposal requires one additional backbone query. In Figure 2, r_t denotes this response and b_t^{+} the clean-motion prediction after generative modulation of the style proposal.

3.5 Backbone Reuse

Flow-matching and rectified-flow backbones (Lipman et al., 2023; Liu et al., 2023) and diffusion or score-based backbones (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b;a) expose different native predictions, but the three priors act on the same clean-motion edits δ_t^k. Under standard VP diffusion notation, this relationship becomes

$$\begin{aligned} b_t = \hat{x}_0(s_B) &= \frac{x_t + \sigma_t^2 s_B(x_t, t, c)}{\alpha_t}, \\ \hat{x}_0(s_B + \Delta s_t^k) = b_t + \delta_t^k \quad &\Longleftrightarrow \quad \Delta s_t^k = \frac{\alpha_t}{\sigma_t^2}\delta_t^k. \end{aligned} \tag{7}$$

Here s_B is the frozen score model, while α_t and σ_t are the clean-signal and noise coefficients. Thus, the same clean-motion edit δ_t^k becomes a velocity correction for flow models and a score correction Δs_t^k for diffusion models. Kimodo predicts $\hat{x}_0$ directly and therefore applies δ_t^k in clean-motion space. Across backbones, we reuse the learned priors and real-motion statistics, changing only representation conversion and sampling strengths while keeping the generator frozen. Lightweight post-processing handles residual motion and contact artifacts after sampling. Training, optimization, representation conversion, backbone adaptation, and post-processing details are consolidated in the supplementary material for reproducibility.

4 Experiments

4.1 Experimental Setup

Benchmark and protocol. We evaluate on the HumanML3D test split, using its descriptions as content text and 100STYLE motions as style references. Text-based controls use MLD and MDM (Chen et al., 2023; Tevet et al., 2023) with manually expanded style descriptions. HY + Qwen3 detail uses Qwen3 (Qwen Team, 2025) to expand each style name into a detailed text prompt. All methods share one task manifest and are grouped by style-text or style-motion input. We omit StyleMotif, MulSMo, and SMooGPT (Guo et al., 2025b; Li et al., 2024; Zhong et al., 2025) because implementations and checkpoints required by our protocol are unavailable, rather than copy scores from incompatible protocols. Main comparisons average three generation seeds; ablations use one fixed seed. The supplementary material consolidates implementation details, evaluator definitions, and per-seed statistics for reproducibility.

Table 1: Main comparison on HumanML3D over three generation seeds. Both input groups receive content text and are ranked separately. HY denotes HY-Motion; LC denotes LongConv. In arrowed metric columns, colors mark the **first**, second, and third values within each input group.

Method	Content alignment			Style recognition					Physical diagnostics		Distribution
	R@1↑	R@3↑	MM↓	SRA-4[a] ↑	Trans.↑	LC↑	TCN↑	BiGRU↑	Skate↓	Slide↓	FID[b] ↓
Style text input											
MLD + manual detail	26.71	51.15	4.93	2.57	2.71	2.50	2.52	2.56	10.57	0.243	1.92
MDM + manual detail	33.02	59.45	4.40	2.28	2.49	2.21	2.22	2.20	10.99	0.293	1.79
HY + short label	**39.48**	67.36	3.94	3.04	3.26	3.18	2.78	2.95	8.27	0.228	**1.10**
HY + manual detail	33.80	59.52	4.52	6.23	6.69	6.84	5.74	5.63	8.34	0.232	1.84
HY + Qwen3 detail	35.95	65.08	4.10	2.92	3.29	3.08	2.78	2.54	7.88	0.230	1.37
Kimodo + short label	28.10	54.70	4.61	2.08	2.15	2.13	1.96	2.07	11.98	0.322	3.24
Kimodo + manual detail	24.54	47.16	5.18	3.18	3.83	3.58	2.57	2.75	12.66	0.339	3.18
HY-Lite + TriPrior	37.28	67.44	3.88	29.26	34.08	27.36	**29.23**	26.35	5.31	0.158	1.85
HY + TriPrior	38.62	**69.39**	**3.80**	29.96	35.66	28.38	29.04	26.76	**4.47**	**0.137**	1.61
Kimodo + TriPrior	27.36	55.44	4.57	**30.81**	**36.07**	**28.77**	28.65	**29.76**	7.76	0.188	3.98
Style motion input											
SMooDi	27.31	55.21	4.58	16.76	17.30	17.14	16.09	16.51	10.17	0.237	**1.53**
MLD + Motion Puzzle	11.23	26.29	6.75	8.95	9.33	10.17	8.06	8.25	18.07	0.376	6.56
MLD + Aberman	27.38	53.35	4.59	13.64	12.58	14.06	14.97	12.96	23.20	0.551	4.60
MoMo	5.94	13.75	8.17	26.34	27.43	26.28	26.49	25.17	27.22	0.487	14.25
PersonaBooth	17.04	35.04	6.11	12.86	11.99	12.31	13.93	13.20	5.50	0.175	9.82
HY-Lite + TriPrior	37.36	67.64	3.88	28.63	33.49	26.76	**28.50**	25.76	5.32	0.159	1.83
HY + TriPrior	**38.70**	**69.50**	**3.79**	29.40	35.05	27.93	28.46	26.18	**4.49**	**0.138**	1.59
Kimodo + TriPrior	27.41	55.35	4.56	**30.30**	**35.50**	**28.45**	28.02	**29.23**	7.75	0.188	3.98

[a]SRA-4 is the arithmetic mean of Transformer (Trans.), LongConv (LC), TCN, and BiGRU accuracy. [b]FID compares with the original HumanML3D content distribution; see the discussion in Section 4.2. Slide is in m/(motion·s); MM and FID are unscaled, and all other metrics are percentages throughout the main comparison.

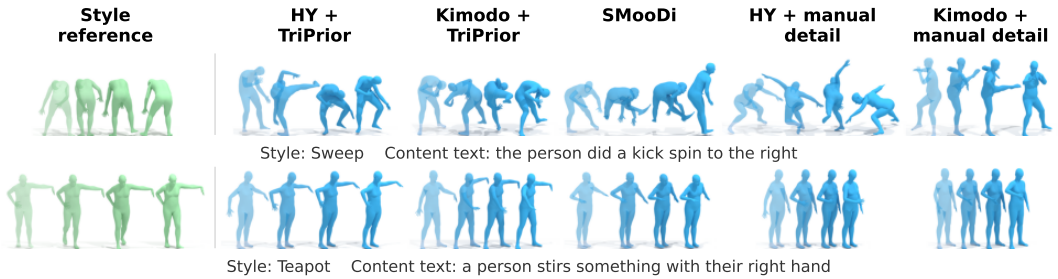

Figure 3: Qualitative comparison of the Table 1 methods under a shared style and seed. Green denotes the 100STYLE reference and blue the stylized outputs; four poses progress from light to dark and are selected at each method's highest leg lift.

Evaluation metrics. R@1, R@3, and multimodal distance (MM-Dist, abbreviated MM) measure content alignment between text and motion. Skate and contact sliding (Slide) diagnose contact quality. FID (Heusel et al., 2017) measures similarity to the original HumanML3D distribution. Because a target style intentionally changes motion statistics, an FID increase may reflect either successful stylization or motion artifacts. We therefore interpret it jointly with content, style, and physical metrics. SRA-4 averages four style classifiers: Transformer (Vaswani et al., 2017), LongConv (Poli et al., 2023), TCN (Bai et al., 2018), and BiGRU (Cho et al., 2014; Schuster & Paliwal, 1997). All methods use the same frozen checkpoints, independent of the generation-guidance networks. We report heterogeneous evaluators because SMooDi's native pipeline (Zhong et al., 2024) reuses its guidance network for scoring, which may reward evaluator-specific features.

4.2 Main Comparison

Overall performance and backbone reuse. Table 1 shows that TriPrior combines strong style recognition with content alignment and retains style control across backbones. On HY with style-motion input, it achieves 29.40% SRA-4 and 38.70% R@1, compared with MoMo's 26.34% and 5.94%. MoMo transfers attention features between two motions (Raab et al., 2024) without explicitly verifying the requested text. In contrast, our content-relative edits and preservation checks retain substantially more content alignment. The same learned priors transfer without retraining to HY-Lite and Kimodo through backbone-specific representation conversion and sampling settings. HY-Lite

Table 2: HY-Motion ablations under the revised frozen configuration, using style-motion input and one seed. SRA-4, R@1, and Skate are percentages; Slide is in m/(motion·s).

Variant	SRA-4↑	FID↓	R@1↑	Skate↓	Slide↓
Full method	29.18	1.606	38.68	4.56	0.1384
w/o dataset prior	22.49	1.403	40.06	5.16	0.1511
w/o instance prior	4.06	0.664	44.11	6.44	0.1907
w/o generative prior	32.20	2.115	37.13	5.56	0.1596
w/o post-processing	29.22	1.813	38.11	8.73	0.2432

Table 5: User study on HY-Motion. Values are judgments favoring TriPrior over the first-column baseline; 50% is parity, ties count as half, and values are rounded to the nearest percentage point for compact reporting.

	Preference for TriPrior (%)↑			
Compared against	Content	Style	Natural.	Overall
SMooDi	65%	71%	69%	74%
HY + manual detail	60%	77%	61%	65%



Figure 4: Qualitative prior ablation with style-motion input and generation seed 42. Green is the 100STYLE reference and blue is the stylized output; frames progress from light to dark. All columns share uniformly spaced frames, one at the largest change without the generative prior.

obtains 28.63% SRA-4 and 37.36% R@1. Kimodo reaches the highest SRA-4 with both style-text (30.81%) and style-motion input (30.30%), although its style-motion R@1 is 27.41%. Thus, shared style control remains effective across generators, while content performance remains backbone dependent. Figure 1 further shows style-dependent changes in arm posture, trunk lean, and support across four contents. Inference requires one cached content rollout and additional late-stage backbone queries; Appendix Table 11 reports NFE, latency, and peak memory.

4.3 Prior and Post-Processing Ablations

Table 2 ablates the three priors under the same frozen HY configuration and seed, without retuning individual variants. The results support the coarse-to-fine design in Section 3. Removing the dataset-level prior lowers SRA-4 from 29.18% to 22.49%, showing that its content-balanced direction provides a stable style initialization. Removing the instance-level prior collapses SRA-4 to 4.06%, indicating that an averaged direction alone cannot supply the motion-specific refinement needed across contents. Both removals raise R@1 and lower FID, consistent with the data priors deliberately moving the content trajectory toward the target style.

The generative prior plays a different role: it modulates the style proposal using the frozen generator rather than establishing style itself. Removing it raises SRA-4 to 32.20% but worsens FID, content retrieval, skating, and sliding. Retaining it therefore shifts the result toward content and motion quality at a measurable cost in style recognition, matching the proposal-modulation mechanism in Section 3.4. Post-processing is complementary: removing it leaves SRA-4 nearly unchanged but sharply worsens Skate and Slide, confirming its role in repairing residual contact artifacts. Together, data priors establish style and generator response modulates the proposal, while post-processing repairs residual artifacts before the final stylized motion is returned.

4.4 User Study

Human judgments favor TriPrior across style, content, and perceptual quality. In a blinded pairwise study, more than 50 participants compared TriPrior on HY-Motion with SMooDi and HY + manual detail on content, style, naturalness, and overall preference. TriPrior exceeds parity in every category (Table 5), including 74% overall and 69% naturalness preference against SMooDi. Against HY + manual detail, the strongest preference is style at 77%, supporting motion-based style conditioning over a written description on the same backbone. Crucially, stronger perceived style does not come at the expense of content or naturalness, whose preferences also remain above parity against both baselines. These judgments complement the automated style, retrieval, and physical metrics, supporting the intended balance among the three priors. Appendix F gives the full protocol.

Table 3: HY generative prior by output stage (SRA-4, R@1, and Skate in %). Terminal is after calibrated response and before selection and repair; on/off pairs share tasks and seed.

Output	Gen. prior	SRA-4↑	R@1↑	Skate↓	FID↓
Sampler	Off	31.42	37.70	10.82	2.342
Sampler	On	30.69	37.49	9.41	2.089
Terminal	Off	32.21	37.42	11.18	2.314
Terminal	On	29.73	38.16	10.55	1.790
Final output	Off	32.20	37.13	5.56	2.115
Final output	On	29.18	38.68	4.56	1.606

Table 4: Interaction between instance and generative priors on final HY outputs. Tasks, seed, dataset prior, and post-processing are fixed; metric units follow Table 3.

Inst. prior	Gen. prior	SRA-4↑	R@1↑	Skate↓	FID↓
Off	Off	4.02	43.99	7.55	0.673
Off	On	4.06	44.11	6.44	0.664
On	Off	32.20	37.13	5.56	2.115
On	On	29.18	38.68	4.56	1.606

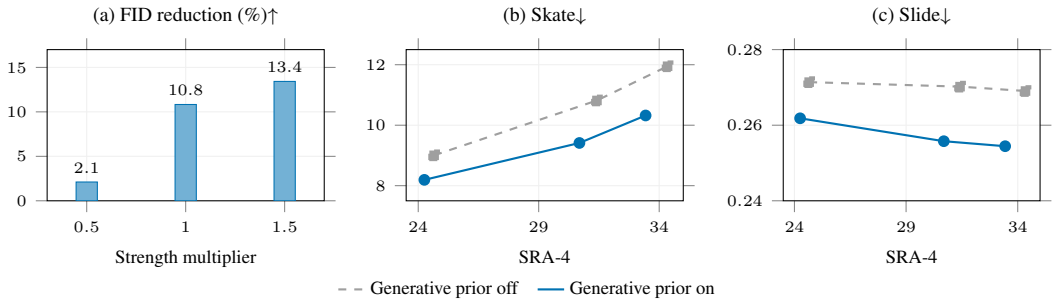

Figure 5: Current frozen HY strength sweep before post-processing, with matched tasks and seed. Panel (a) reports paired relative FID reduction; panels (b,c) use achieved SRA-4 on the horizontal axis. SRA-4 and Skate are percentages, and Slide is in m/(motion·s).

4.5 Quality–Style Trade-off of the Generative Prior

Table 3 separates the generative prior from later repair. FID and Skate improve already at the sampler output, showing that the generator response changes the sampling result rather than relying on post-processing. After calibration, the content benefit also emerges and persists to the final output: FID improves from 2.115 to 1.606 and R@1 from 37.13% to 38.68%. SRA-4 simultaneously decreases from 32.20% to 29.18%. This matches Section 3.4: the content-conditioned response modulates a strong style proposal toward content and motion quality.

Table 4 further links this trade-off to the instance-level proposal. With the instance-level prior active, generator modulation produces substantial FID and R@1 gains together with a 3.02-point SRA-4 cost. Without that proposal, FID, R@1, and SRA-4 remain nearly unchanged. The content and distribution effects therefore arise mainly when the generator modulates a strong style edit. Skate improves in both cases, suggesting a broader benefit to motion regularity across proposal regimes.

Finally, Figure 5 shows paired FID reductions of $2.1/10.8/13.4\%$; Skate and Slide also improve, with less than one point of SRA-4 and R@1 change. Together with Tables 3 and 4, this supports our central claim: the frozen generator acts as an active motion prior, regularizing style edits toward better distributional and physical quality while exposing the style–content–quality trade-off during inference across all evaluated strengths.

5 Conclusion

TriPrior decouples unpaired style control from pretrained motion generators by coordinating two data priors with a generative prior supplied by the frozen backbone. Three-seed results on HY-Motion, together with reuse on HY-Lite and Kimodo, show that the data-prior components transfer through backbone-specific representation conversion without training an internal style branch. Controlled HY-Motion experiments further show that generator response improves FID, content retrieval, and skating immediately after style editing, and that these gains persist after post-processing alongside a measurable style–quality trade-off. Future work should reduce the added inference cost and extend perceptual evaluation across broader styles and content motions.

AI USE STATEMENT

In this work, we used generative AI tools for code inspection, experimental planning, analysis implementation, and language editing. All AI-assisted code and numerical summaries were checked against executable analyses and frozen artifacts, and literature statements were checked against primary sources. The authors reviewed all AI-assisted work and take responsibility for the final text, claims, code, and artifacts.

ETHICS STATEMENT

This study analyzes previously collected motion datasets and includes a perceptual study in which participants compare anonymized rendered motion clips. We report only aggregate preference rates and no identifying participant information. Generated human motion may inherit representation and coverage biases from the underlying datasets, so conclusions are restricted to the reported benchmark and style set.

REPRODUCIBILITY STATEMENT

Section 4.1 specifies task, data, evaluator, and comparison contracts. Section 3.1 defines the stage order and shared edit interface, while Sections 3.2–3.4 define the three priors. Each completed audit is backed by a frozen JSON configuration, machine-readable result, provenance record, and a one-command analysis entry point. Missing primary artifacts and failed converter or reconstruction gates produce explicit unavailable or failed states rather than silent protocol substitutions. An anonymized release will include the registry, hashes, evaluator identities, and scripts needed to rebuild the reported tables from frozen outputs.

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*, 2020. doi: 10.1145/3386569.3392469.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Ziyi Chang, Edmund J. C. Findlay, Haozheng Zhang, and Hubert P. H. Shum. Unifying human motion synthesis and style transfer with denoising diffusion probabilistic models, 2022.

Kerui Chen, Jianrong Zhang, Ming Li, Zhonglong Zheng, and Hehe Fan. ClusterStyle: Modeling intra-style diversity with prototypical clustering for stylized motion generation. In *European Conference on Computer Vision*, volume 17047 of *Lecture Notes in Computer Science*, pp. 599–618. Springer, 2026. doi: 10.1007/978-3-032-37314-4_33.

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

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022.

Tong Guo, Zhiqian Xia, Feng Yu, Haifeng Xia, and Siyu Xia. A generalized two-stage approach to motion style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 6626–6635, 2025a.

Ziyu Guo, Young Yoon Lee, Joseph Liu, Yizhak Ben-Shabat, Victor Zordan, and Mubbasir Kapadia. StyleMotif: Multi-modal motion stylization using style–content cross fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13349–13359, 2025b.

Meenakshi Gupta, Mingyuan Lei, Tat-Jen Cham, and Hwee Kuan Lee. D-LORD: Learning disentangled representations for individual style and motion in human actions, 2024.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, 35(4):138:1–138:11, 2016. doi: 10.1145/2897824.2925975.

Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima, and Taku Komura. Fast neural style transfer for motion data. *IEEE Computer Graphics and Applications*, 37(4):42–49, 2017. doi: 10.1109/MCG.2017.3271464.

Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics*, 2022. doi: 10.1145/3516429.

Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. MoST: Motion style transformer between diverse action contents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1705–1714, 2024.

Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, et al. Personabooth: Personalized text-to-motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zhe Li, Yisheng He, Lei Zhong, Weichao Shen, Qi Zuo, Lingteng Qiu, Zilong Dong, Laurence Tianruo Yang, and Weihao Yuan. MulSMo: Multimodal stylized motion generation by bidirectional control flow, 2024.

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

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

Ian Mason, Sebastian Starke, He Zhang, Hakan Bilen, and Taku Komura. Few-shot learning of homogeneous human locomotion styles. *Computer Graphics Forum*, 37(7):143–153, 2018. doi: 10.1111/cgf.13555.

Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1), May 2022. doi: 10.1145/3522618.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Qwen Team. Qwen3 Technical Report, 2025.

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 *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–13, 2024. doi: 10.1145/3680528.3687579.

Davis Rempe, Mathis Petrovich, Ye Yuan, et al. Kimodo: Scaling controllable human motion generation, 2026.

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

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

Tianxin Tao, Xiaohang Zhan, Zhongquan Chen, and Michiel van de Panne. Style-ERD: Responsive and coherent online motion style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6593–6603, 2022.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Yuxin Wen, Qing Shuai, Di Kang, et al. HY-Motion 1.0: Scaling flow matching models for text-to-motion generation, 2025.

Fatemeh Zargarbashi, Dhruv Agrawal, Jakob Buhmann, Martin Guay, Stelian Coros, and Robert W. Sumner. VQ-Style: Disentangling motion and style for diverse motion generation. *Computer Graphics Forum*, 2026. doi: 10.1111/cgf.70377.

Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. SMooDi: Stylized motion diffusion model. In *European Conference on Computer Vision*, 2024.

Lei Zhong, Yi Yang, and Changjian Li. SMooGPT: Stylized motion generation using large language models, 2025. Pacific Graphics 2025, Computer Graphics Forum journal track.