Retrieving and Refining Winning Noise Tickets for Diffusion-Based Motion Generation

Sakuya Ota[1], Qing Yu[2], Kent Fujiwara[2], Satoshi Ikehata[3], and Ikuro Sato[1]

[1] Institute of Science Tokyo, Tokyo, Japan
[2] LY Corporation, Tokyo, Japan
[3] National Institute of Informatics (NII), Tokyo, Japan
ota_sakuya@d-itlab.c.titech.ac.jp

Abstract. Diffusion-based text-to-motion models synthesize realistic human motions but often exhibit semantic drift from the input text. Motion is inherently temporal, especially in compositional and long-duration sequences that require semantic consistency across multiple action segments and smooth kinematic transitions throughout the trajectory. We posit that the initial noise is central to this consistency: within the Gaussian noise space, certain instances, *i.e. winning noise tickets*, carry latent structure that biases denoising toward particular motion semantics, even under null prompts. We propose WInning Noise Retrieval and Optimization (WINRO), a training-free, model-agnostic framework that improves text–motion alignment by selecting and refining such tickets before diffusion sampling. WINRO maps random noises to motion features generated under null prompts, retrieves the best-aligned noise for a given text, and refines it via a KL-regularized objective that reduces the residual semantic gap while preserving the Gaussian prior. An optional LoRA-based adapter amortizes this refinement into a single forward pass. WINRO consistently improves text–motion fidelity across different base models, MDM and MotionLCM, on HumanML3D without retraining, improves temporal robustness on the MTT benchmark, and generalizes to applications such as motion stylization and spatial constraint satisfaction.

Keywords: Human motion generation · Diffusion models · Text-to-motion · Noise optimization

1 Introduction

Human motion generation is a fundamental task in computer graphics, robotics, and virtual reality, where the goal is to synthesize realistic, semantically coherent motion sequences from high-level inputs. In particular, text-to-motion (T2M) diffusion models translate natural language descriptions into motion trajectories, enabling intuitive workflows for animation and embodied AI systems [1, 3, 9, 16, 28, 31, 39]. Despite their success, diffusion-based T2M models often suffer from semantic drift, a challenge particularly pronounced in temporally structured, long-horizon generation, where the text specifies multiple action segments and

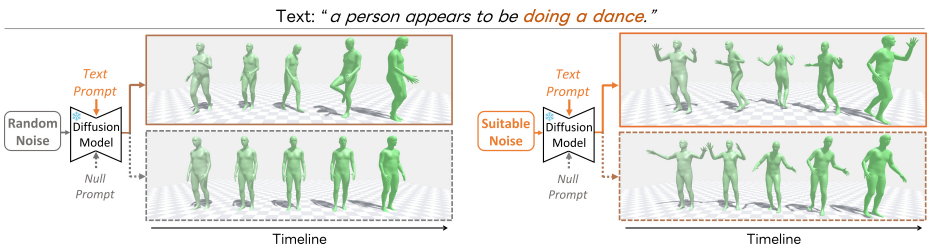


Fig. 1: Winning Noise Ticket Hypothesis for motion. Certain initial noises inherently encode motion semantics, producing meaningful motion even under a null prompt (right), while others yield static poses regardless of conditioning (left). WINRO retrieves and refines such tickets to improve text–motion alignment.

the generated trajectory must realize them with consistent semantics and smooth kinematic transitions. This limits their applicability in scenarios demanding fine-grained control over style or body-part behavior.

We posit that within the Gaussian noise space used to initialize diffusion-based motion generators, there exist specific noise instances, termed *winning noise tickets*, that consistently bias the denoising trajectory toward temporally coherent motions with particular semantics or styles. This notion builds on the extension of the Lottery Ticket Hypothesis from network pruning [7] to diffusion models [22]. Empirically, certain noise samples produce coherent, meaningful motions even under null prompts, while others yield implausible or text-misaligned results even when conditioned on descriptive text (Fig. 1). Thus, the initial noise is not mere randomness, but acts as a latent semantic prior that shapes the generative trajectory. While similar observations have been made for image diffusion [22], the effect is especially impactful in motion generation, where training data is orders of magnitude scarcer (approximately 15k samples vs. billions of images), making models more sensitive to the choice of initial noise.

Prior works that optimize initial noise in motion diffusion do so to satisfy physical or geometric constraints [17, 20]. Because their objectives are defined over predefined spatial targets rather than the text prompt itself, the optimized noise may fulfill the physical goal while still producing motion that drifts from the intended textual meaning. We address this gap by directly optimizing noise against text–motion semantic similarity, and ask: *Can we identify and refine the winning noise tickets that best embody the semantics of a given text prompt?*

To address this, we introduce **WI**nning **N**oise **R**etrieval and **O**ptimization (**WINRO**), a *training-free and model-agnostic* framework that improves text–motion alignment by discovering and refining winning noise tickets before the diffusion process. Selecting a semantically aligned noise requires associating each noise sample with a descriptor that (i) captures motion semantics and (ii) is comparable with text. We achieve this by constructing a *noise dictionary*. We sample a random Gaussian noise and use a frozen diffusion model to generate a motion under a null prompt. A pretrained text–motion retrieval model then encodes the motion into a feature vector in a shared text–motion embedding space, which we store with the noise as a dictionary entry. Given an input text prompt,

WINRO retrieves the noise whose motion feature is closest to the text embedding in this space, providing a semantically favorable initialization. The retrieved noise is then refined through a *KL-regularized objective* that minimizes the residual text–motion discrepancy while preserving the Gaussian prior. Optionally, a trained LoRA-based Noise Refiner can amortize this iterative refinement into a single forward pass. The key contributions are:

- We identify the *Winning Noise Ticket* in motion diffusion, showing that certain initial noises inherently bias denoising trajectories toward specific motion semantics.
- We propose WINRO, a training-free framework for *motion-specific* semantic noise retrieval and refinement, featuring a null-prompt motion dictionary, frame-adjusted retrieval across variable durations, and KL-regularized optimization in the text–motion embedding space.
- We optionally extend WINRO with a LoRA-based Noise Refiner that, once trained, amortizes test-time optimization into a single forward pass without updating the base model weights.
- We demonstrate consistent improvements in text–motion fidelity across diffusion backbones on HumanML3D, improved long-horizon temporal robustness on the MTT benchmark, and generalization to motion stylization and spatial constraint satisfaction.

2 Related Work

Motion Generation Models. Advances in motion-capture and video-based reconstruction [18, 29, 35] have driven interest in generating realistic human motion. Early methods focused on motion prediction or interpolation [12, 38], later extending to action-conditioned synthesis [25, 34]. More recently, *text-to-motion (T2M)* diffusion models [9, 10, 15, 39] have emerged as powerful tools for natural-language-driven motion generation. Notably, MotionDiffuse [40], MDM [31], MotionLCM [4], and the skeleton-aware latent diffusion model SALAD [13] achieve strong realism but treat initial noise as an unstructured random seed, leaving its semantic role unexplored.

Motion-Text Retrieval. Motion-text retrieval models [8, 26, 36] learn aligned multimodal spaces for semantic comparison. Retrieval-augmented generation (RAG) frameworks such as ReMoDiffuse [41] retrieve motion exemplars to condition the denoising process, improving fidelity through external references [19, 37]. Rather than retrieving exemplar motions, our method retrieves semantically meaningful initial noises whose induced null-prompt motions align with the input text, making it complementary to exemplar-based retrieval.

Initial Noise Selection and Optimization. Recent studies show that the initial noise in diffusion models affects generated outcomes [21, 33], motivating noise optimization [2, 5, 6, 11] and selection [22, 32]. The *Lottery Ticket Hypothesis in Denoising* [22] suggests that certain "winning" noise patterns bias denoising

toward specific semantics. In motion generation, such biases are consequential, as the initial noise can affect temporal semantic consistency. Accordingly, in the motion domain, works such as DNO [17], ProgMoGen [20], and PINO [24] optimize noise via physical or geometric constraints. In this research, we target text–motion semantic alignment, retrieving noise from a precomputed dictionary and refining it against the text prompt.

3 Winning Noise Tickets for Motion Diffusion

3.1 Diffusion-Based Motion Generation

We build upon diffusion-based text-to-motion (T2M) models, which generate motion by progressively denoising Gaussian noise. A motion sequence $\mathbf{x} \in \mathbb{R}^{F \times D}$ has F frames and D=263 features [10]. The forward process gradually corrupts clean motion $\mathbf{x}_0$ into noise $\mathbf{x}_T$ over T steps:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{1}$$

where α_t is the noise schedule. With a sufficiently long schedule, we approximate $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The denoising network G_θ learns to invert this process:

$$\mathbf{x}_0 = G_\theta(\mathbf{x}_T, c) = [f_0 \circ f_1 \cdots \circ f_T](\mathbf{x}_T, c), \tag{2}$$

where f_t is the t-th denoising update and c is a conditioning input such as text. Models like MDM [31] operate in the motion domain, while latent variants (*e.g.*, MLD [3]) denoise compact motion latents. With deterministic samplers (*e.g.*, DDIM [30]), there is a one-to-one mapping between the initial noise $\mathbf{x}_T$ and the generated motion $\mathbf{x}_0$, making the initial noise a controllable factor.

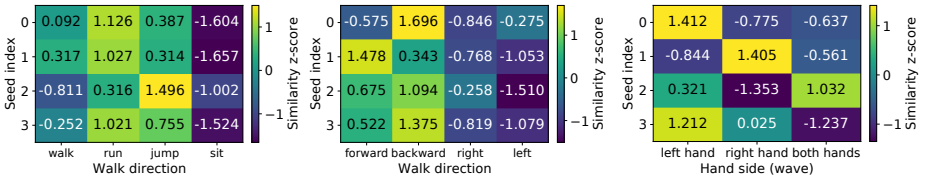
3.2 Theoretical Perspective: Why Noise Encodes Motion Semantics

Under deterministic sampling, diffusion defines a continuous map from initial noise to a motion trajectory, so nearby noises tend to produce similar motions. Composed with a retrieval encoder, the resulting text–motion alignment score varies smoothly over the noise space; hence high-alignment initial noise form *regions* rather than isolated points, enabling both retrieval and local refinement.

Setup. Let $\mathcal{F}_{\text{Motion}}$ and $\mathcal{F}_{\text{Text}}$ be the motion–text retrieval encoders. For deterministic sampling, define the cosine alignment score

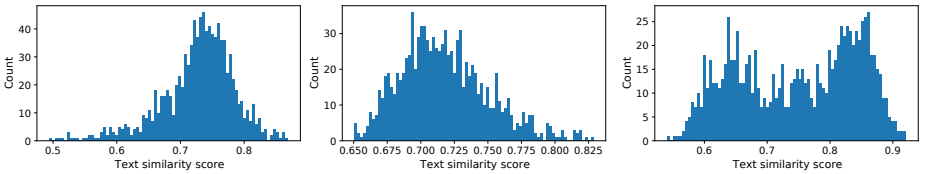
$$S_c(\mathbf{x}_T) = \frac{\mathcal{F}_{\text{Motion}}(G_\theta(\mathbf{x}_T, c))^\top \mathcal{F}_{\text{Text}}(c)}{\|\mathcal{F}_{\text{Motion}}(G_\theta(\mathbf{x}_T, c))\| \, \|\mathcal{F}_{\text{Text}}(c)\|} \in [-1, 1]. \tag{3}$$

We use the *null* score $\tilde{S}_c(\mathbf{x}_T)$ to describe intrinsic semantics under unconditional generation, defined by setting the diffusion condition to c=$\varnothing$ while keeping the query text embedding $\mathcal{F}_{\text{Text}}(c)$. We call $\mathbf{x}_T$ a *winning noise ticket* for c (level γ) if $\tilde{S}_c(\mathbf{x}_T) \geq \gamma$.



(a) Action category **(b)** Action direction **(c)** Action parts

Fig. 2: z-score heatmaps of action-specific tendencies under four fixed noise seeds. For each seed and prompt group, we generate ten motions with varied prompts, compute the mean TMR similarity per category, and apply per-seed z-score normalization.



(a) "walk" **(b)** "run" **(c)** "jump"

Fig. 3: Similarity-score histograms for the fixed prompts "the person walks/runs/jumps." For each prompt, we compute TMR similarity scores over 1,000 initial noise seeds.

Smoothness implies semantic regions. If the deterministic sampler and $\mathcal{F}_{\text{Motion}}$ are locally Lipschitz, then $\tilde{S}_c$ is locally Lipschitz in $\mathbf{x}_T$. Therefore, whenever $\tilde{S}_c(\mathbf{x}_T^\star) \geq \gamma$, there exists a radius $r > 0$ such that all $\mathbf{x}_T$ with $\|\mathbf{x}_T - \mathbf{x}_T^\star\| \leq r$ also satisfy $\tilde{S}_c(\mathbf{x}_T) \geq \gamma/2$. This locality motivates (i) best-of-N dictionary retrieval and (ii) small, prior-preserving refinements (Sec. 4).

3.3 Evidence of Winning Noise Tickets

To examine whether initial noise influences motion semantics, we analyze T2M generation using MDM [31] and a text–motion retrieval model TMR [26], measuring text–motion similarity as cosine similarity in the shared embedding space.

Bias Patterns Under Fixed Noise. We fix four random noise seeds and measure text–motion similarity across prompt groups targeting action categories, directions, and body parts. Fig. 2 shows clear seed-specific preferences, especially for direction and body-part movement (Fig. 2b, 2c), indicating distinct semantic regions in the noise space.

Alignment Variability Under Noise Variation. Conversely, fixing three prompts and varying the noise over 1,000 seeds reveals prompt-dependent sensitivity (Fig. 3): some prompts have only a small subset of highly compatible seeds, while others show broader variability. This confirms that prompt-specific winning noise tickets exist, motivating retrieval and refinement.

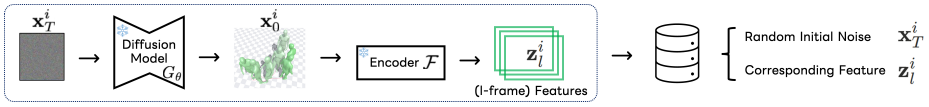


Fig. 4: The pipeline of building noise dictionary $\mathcal{D}$. We randomly sample noise $\mathbf{x}_T^i$ and use a diffusion model G_θ and encoder $\mathcal{F}$ of a retrieval model to generate a feature $\mathbf{z}_l^i$. For the frame-adjusted dictionary, we truncate the output motion $\mathbf{x}_0^i$ at frame l and generate the features.

We build on these observations in Sec. 4 by retrieving promising seeds from a null-prompt dictionary (Sec. 4.1) and refining them with a KL-regularized objective that preserves the Gaussian prior (Sec. 4.2).

4 Winning Noise Retrieval and Optimization

We propose to exploit this semantic bias of initial noise with two core stages: (i) retrieving a semantically favorable noise seed from a *null-prompt noise dictionary* built in a text–motion embedding space, yielding WIN-FAR, a training-free and real-time variant, and (ii) optionally refining that seed by optimizing a KL-regularized text–motion similarity objective, yielding the full WINRO pipeline, which remains training-free but is iterative and therefore suited for offline or quality-critical use. Under deterministic sampling, the alignment varies smoothly with $\mathbf{x}_T$ (Sec. 3.2), so retrieval provides a good initialization for local refinement. Both stages are training-free and model-agnostic.

4.1 Stage 1: Winning Noise Ticket Retrieval

We build a noise dictionary $\mathcal{D}$ using a frozen diffusion model G_θ and a pre-trained motion–text retrieval model (Fig. 4). This construction is a one-time offline cost for each backbone, after which the dictionary is reused across all queries. Per-query inference requires only a retrieval lookup and optional refinement, without additional diffusion runs for dictionary construction. For each entry $i \in [1, \cdots, N]$, we sample $\mathbf{x}_T^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, generate a null-prompt motion $\mathbf{x}_0^i = G_\theta(\mathbf{x}_T^i, \varnothing)$, and store its retrieval feature $\mathbf{z}_l^i = \mathcal{F}_{\text{Motion}}(\mathbf{x}_0^i)$:

$$\mathcal{D}_{reg} = \{(\mathbf{z}_l^i, \mathbf{x}_T^i)\}_{i=1}^N . \tag{4}$$

To support retrieval across varying target durations, we introduce frame-adjusted retrieval (FAR). Specifically, we encode truncated motions at multiple lengths $l' \in [m, M]$ and store their length-specific retrieval features:

$$\mathcal{D}_{adj} = \{(\mathbf{z}_{l'}^i, \mathbf{x}_T^i) \mid i \in [1, N],\ l' \in [m, M]\} . \tag{5}$$

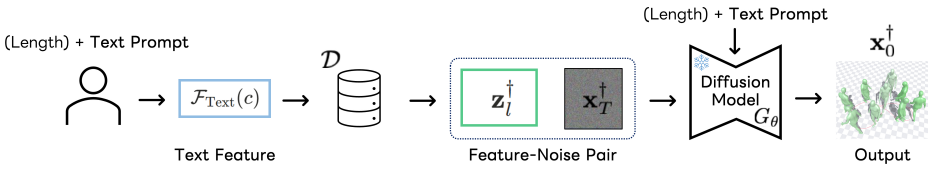


Fig. 5: Overview of Winning Noise Ticket Retrieval. The input text c and, optionally, the desired motion length l are used to retrieve the closest feature $\mathbf{z}_l^\dagger$, whose corresponding noise $\mathbf{x}_T^\dagger$ is then used to generate the desired motion. Note that we freeze the feature encoder and diffusion model.

Why retrieval works. Let $\tilde{S}_c(\mathbf{x}_T)$ denote the null-prompt alignment score used for retrieval (Sec. 3.2). If $p_{c,\gamma} = \mathbb{P}(\tilde{S}_c(\mathbf{x}_T) \geq \gamma)$, then for N independent samples, $\mathbb{P}(\max_{i\in[N]} \tilde{S}_c(\mathbf{x}_T^{(i)}) \geq \gamma) = 1 - (1 - p_{c,\gamma})^N$. Thus, the probability of sampling a winning region increases with dictionary size N, but with diminishing returns. Moreover, null-prompt retrieval can transfer to text prompts when conditioning acts as a limited perturbation in the retrieval space, consistent with classifier-free guidance mixing unconditional and conditional predictions. The detailed proof is provided in the supplementary materials.

At retrieval time, we compute the cosine similarity between each dictionary feature and the text embedding:

$$s(\mathbf{z}, c) = \frac{\mathbf{z} \cdot \mathcal{F}_{\text{Text}}(c)}{\|\mathbf{z}\| \|\mathcal{F}_{\text{Text}}(c)\|}, \tag{6}$$

where $\mathcal{F}_{\text{Text}}$ is the text encoder of the retrieval model. We rank all entries by similarity and define the top-k candidate set:

$$\mathcal{N}_k(c) = \underset{(\mathbf{z}_l^i, \mathbf{x}_T^i)\in\mathcal{D}}{\text{Top-k}} \; s(\mathbf{z}_l^i, c). \tag{7}$$

where $\mathcal{D} \in \{\mathcal{D}_{\text{reg}}, \mathcal{D}_{\text{adj}}\}$. The retrieved feature $\mathbf{z}^\dagger$ is then sampled uniformly from $\mathcal{N}_k(c)$. When k=1, this reduces to deterministic argmax retrieval. Along with the provided text prompt c, the initial noise $\mathbf{x}_T^\dagger$ corresponding to $\mathbf{z}^\dagger$ is used to generate the final motion (Fig. 5). We denote retrieval using $\mathcal{D}_{reg}$ as "Retrieval" and using $\mathcal{D}_{adj}$ as "Frame-adjusted retrieval (FAR)".

Noise perturbation. To avoid over-concentrating on a small subset of seeds while preserving semantics, we perturb the retrieved noise with small Gaussian jitter:

$$\mathbf{x}'_T = \mathbf{x}_T^\dagger + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{8}$$

where σ controls the stochasticity.

4.2 Stage 2: Winning Noise Ticket Refinement

We refine the retrieved noise $\mathbf{x}_T^\dagger$ by maximizing text–motion similarity (Fig. 6):

$$\mathcal{L}_{sim} = 1 - s(\mathcal{F}_{\text{Motion}}(\mathbf{x}_0), c). \tag{9}$$

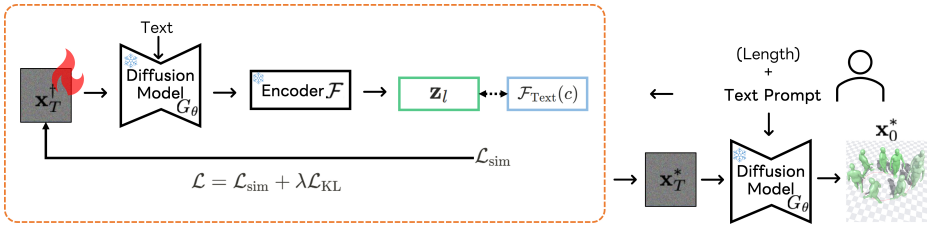


Fig. 6: Overview of Winning Noise Ticket Refinement. The retrieved noise $\mathbf{x}_T^\dagger$ can further be refined to improve the quality of the output using input text c, and optionally the desired motion length l. The loss $\mathcal{L}$ is minimized to obtain the refined noise $\mathbf{x}_T^*$.

As in retrieval, $\mathbf{x}_0$ can be truncated to the target length before encoding. Directly optimizing this similarity loss can move the noise outside the Gaussian typical set (e.g., by drifting its empirical mean/variance), where the diffusion model is poorly calibrated and motion quality/diversity may degrade. We therefore add a KL-divergence term that penalizes deviations from the standard normal:

$$\mathcal{L}_{KL} = D_{KL}\big(\mathcal{N}(\mu(\mathbf{x}_T), \sigma^2(\mathbf{x}_T)) \| \mathcal{N}(0,1)\big)\ , \tag{10}$$

where $\mu(\mathbf{x}_T)$ and $\sigma^2(\mathbf{x}_T)$ are the empirical mean and variance of the noise tensor. The full objective is:

$$\mathcal{L} = \mathcal{L}_{sim} + \lambda\mathcal{L}_{KL}\,, \tag{11}$$

where λ is the balancing parameter. We freeze the motion diffusion model and optimize only the initial noise to obtain $\mathbf{x}_T^*$ aligned with the user-provided text.

4.3 LoRA-Based Noise Refiner

Iterative refinement runs diffusion sampling at every optimization step, which is costly (see AITS in Table 2). As an amortized alternative, we train a Low-Rank Adaptation (LoRA) [14] module that predicts a noise correction in a single forward pass, following [5]. Unlike the retrieval and optimization stages, the LoRA refiner requires an additional training step and is therefore not training-free, but it enables real-time inference:

$$\hat{\mathbf{x}}_T = \mathbf{x}_T + R_\phi(\mathbf{x}_T)\,. \tag{12}$$

Here, R_ϕ predicts a residual correction so that $\hat{\mathbf{x}}_T$ approximates the refined noise produced by retrieval and iterative optimization. Only the LoRA weights ϕ are updated while all diffusion parameters θ remain frozen. During training, $\mathbf{x}_T$ is sampled from the standard Gaussian prior rather than from the retrieval dictionary, so no explicit noise-to-noise supervision pairs are required. The refiner minimizes a loss analogous to Eq. 11:

$$\mathcal{L}_{LoRA} = 1 - \cos(\mathcal{F}_{\text{Motion}}(G_\theta(\hat{\mathbf{x}}_T, c)), \mathcal{F}_{\text{Text}}(c)) + \lambda\|R_\phi(\mathbf{x}_T)\|_2^2, \tag{13}$$

where the second term regularizes the correction magnitude, helping the corrected noise maintain proximity to the Gaussian prior. At inference, the refiner can be applied either to random noise or to the retrieved $\mathbf{x}_T^\dagger$ from Stage 1.

5 Experiments

In this section, we first verify the intrinsic semantics of winning noise tickets through null-prompt generation (Sec. 5.1), then evaluate the full WINRO pipeline on text-conditioned benchmarks (Sec. 5.2). We further examine compositional and long-duration generation (Sec. 5.3) and demonstrate task adaptability to motion stylization and controllable generation (Sec. 5.4). Experimental details are provided in the supplementary material (Sec. B).

5.1 Intrinsic Semantics of Winning Noise Tickets

Generation with Null Prompt. To isolate the effect of noise selection from text conditioning, we first retrieve winning noise tickets using text embeddings but generate motions under a *null prompt*, so that any observed semantic alignment must originate from the noise itself. Following Section 4.1, we construct a noise dictionary containing 10,000 randomly sampled noise–motion pairs, where each feature is computed by encoding a null-prompt motion. For efficiency, truncated samples are created every 4 frames, resulting in motions of lengths $40, 44, \ldots, 196$ for Frame-Adjusted Retrieval (FAR). All noises are drawn from a standard Gaussian distribution, and no ground-truth motion data is used.

We conduct experiments on the HumanML3D dataset [10]. Two representative diffusion backbones are adopted: MDM-50step [31] and MotionLCM [4]. The pretrained TMR model [26] is used to embed both motions and texts into a shared 256-dimensional feature space for retrieval. Unless otherwise noted, we use top-1 retrieval (k=1). Details are provided in the supplementary material.

Following prior works [3, 31], we report FID, R-Precision, Diversity, Multimodal Distance (MMDist), MultiModality, and additionally include AITS. FID measures the distributional distance between generated and real motions in a learned feature space, reflecting overall motion quality. R-Precision and MMDist evaluate text–motion alignment, while Diversity measures variation across generated motions. MultiModality measures variation among motions generated from the same text; see Section C. **AITS (Average Inference Time per Sentence)** reports batch-size-one wall-clock inference time in seconds, excluding data/model loading and the one-time noise dictionary construction. All reported experiments use a single NVIDIA H100 GPU, except where explicitly stated.

As shown in Table 1, retrieved winning noise tickets enable both MDM-50step and MotionLCM to produce semantically aligned motions even under null prompts, where baselines yield near-zero alignment. This validates the Winning Noise Ticket Hypothesis (Sec. 3) by showing that semantic information can be partially encoded in the initial noise itself, independent of text conditioning.

5.2 Generation with Refined Winning Noise Tickets

We next examine whether this effect carries over to text-conditioned generation with the full WINRO pipeline. The retrieved noise $\mathbf{x}_T^\dagger$ from the FAR dictionary is optimized using the KL-regularized objective (Eq. 11), yielding a refined

Table 1: Performance of MDM-50step and MotionLCM using only WIN retrieval on HumanML3D under generation with *null* prompts. Results are averaged over 20 evaluation runs with 95% confidence intervals.

Method	FID ↓	R-Top1 ↑	R-Top2 ↑	R-Top3 ↑	Diversity →	MMDist ↓
GT	$0.002^{\pm.000}$	$0.511^{\pm.002}$	$0.704^{\pm.003}$	$0.798^{\pm.002}$	$9.458^{\pm.088}$	$2.970^{\pm.007}$
MDM-50step	$4.956^{\pm.150}$	$0.037^{\pm.002}$	$0.070^{\pm.002}$	$0.102^{\pm.004}$	$8.101^{\pm.087}$	$8.932^{\pm.039}$
w/ WIN-R	$4.500^{\pm.228}$	$0.250^{\pm.006}$	$0.385^{\pm.009}$	$0.474^{\pm.008}$	$8.482^{\pm.068}$	$5.540^{\pm.057}$
w/ WIN-FAR	$\mathbf{4.458}^{\pm.200}$	$\mathbf{0.274}^{\pm.006}$	$\mathbf{0.415}^{\pm.006}$	$\mathbf{0.507}^{\pm.007}$	$\mathbf{8.547}^{\pm.059}$	$\mathbf{5.325}^{\pm.046}$
MotionLCM	$8.460^{\pm.072}$	$0.034^{\pm.001}$	$0.067^{\pm.002}$	$0.099^{\pm.002}$	$6.646^{\pm.064}$	$8.313^{\pm.010}$
w/ WIN-R	$1.681^{\pm.071}$	$0.404^{\pm.002}$	$0.597^{\pm.004}$	$0.710^{\pm.003}$	$8.386^{\pm.029}$	$3.613^{\pm.020}$
w/ WIN-FAR	$\mathbf{1.569}^{\pm.071}$	$\mathbf{0.424}^{\pm.003}$	$\mathbf{0.617}^{\pm.004}$	$\mathbf{0.728}^{\pm.004}$	$\mathbf{8.424}^{\pm.039}$	$\mathbf{3.485}^{\pm.021}$

Table 2: Performance of MDM-50step and MotionLCM with WINRO on HumanML3D under text-conditioned generation. MotionLCM uses 1-step inference. Results are averaged over 20 evaluation runs with 95% confidence intervals.

Method	FID ↓	R-Top1 ↑	R-Top2 ↑	R-Top3 ↑	Div. →	MMDist ↓	AITS ↓
GT	$0.002^{\pm.000}$	$0.511^{\pm.002}$	$0.704^{\pm.003}$	$0.798^{\pm.002}$	$9.458^{\pm.088}$	$2.970^{\pm.007}$	-
MDM-50step	$0.438^{\pm.009}$	$0.469^{\pm.002}$	$0.661^{\pm.002}$	$0.763^{\pm.002}$	$10.025^{\pm.066}$	$3.258^{\pm.009}$	**0.265**
w/ WIN-R	$0.250^{\pm.007}$	$0.505^{\pm.002}$	$0.702^{\pm.003}$	$0.799^{\pm.003}$	$10.303^{\pm.048}$	$3.047^{\pm.012}$	–
w/ WIN-FAR	$0.248^{\pm.009}$	$0.512^{\pm.003}$	$0.706^{\pm.003}$	$0.804^{\pm.002}$	$10.295^{\pm.033}$	$3.013^{\pm.011}$	0.268
w/ WINRO	$0.115^{\pm.002}$	$0.539^{\pm.002}$	$0.727^{\pm.004}$	$0.817^{\pm.002}$	$9.657^{\pm.026}$	$2.883^{\pm.010}$	81.408
w/ WIN-RLoRA	$\mathbf{0.102}^{\pm.002}$	$\mathbf{0.546}^{\pm.003}$	$\mathbf{0.741}^{\pm.003}$	$\mathbf{0.832}^{\pm.002}$	$\mathbf{9.577}^{\pm.025}$	$\mathbf{2.846}^{\pm.007}$	0.280
MotionLCM	$0.093^{\pm.004}$	$0.548^{\pm.003}$	$0.741^{\pm.003}$	$0.835^{\pm.002}$	$9.566^{\pm.069}$	$2.763^{\pm.008}$	**0.029**
w/ WIN-R	$0.091^{\pm.004}$	$0.566^{\pm.003}$	$0.762^{\pm.002}$	$0.852^{\pm.002}$	$9.424^{\pm.028}$	$2.699^{\pm.007}$	–
w/ WIN-FAR	$0.083^{\pm.003}$	$0.572^{\pm.003}$	$0.767^{\pm.002}$	$0.855^{\pm.001}$	$\mathbf{9.474}^{\pm.028}$	$2.674^{\pm.006}$	0.033
w/ WINRO	$\mathbf{0.072}^{\pm.002}$	$\mathbf{0.580}^{\pm.002}$	$\mathbf{0.775}^{\pm.002}$	$\mathbf{0.861}^{\pm.001}$	$9.242^{\pm.026}$	$\mathbf{2.609}^{\pm.005}$	6.870
w/ WIN-RLoRA	$0.083^{\pm.002}$	$\mathbf{0.580}^{\pm.002}$	$0.772^{\pm.001}$	$0.860^{\pm.001}$	$9.550^{\pm.032}$	$2.649^{\pm.005}$	0.043

noise $\mathbf{x}_T^*$. The optimization is performed at inference time without modifying model parameters. Details are in the supplementary material. Alternatively, a *LoRA Noise Refiner* can approximate this refinement in a single feed-forward pass; when combined with retrieval (WIN-RLoRA), it amortizes the iterative optimization while preserving the benefit of a semantically aligned initialization.

Table 2 summarizes the text-conditioned generation results. Each stage of WINRO yields consistent improvements across both diffusion backbones. Plain retrieval (WIN-R) already improves text alignment over the baselines, reducing the FID for MDM-50step from 0.438 to 0.250. Frame-adjusted retrieval with perturbation (WIN-FAR) further improves upon WIN-R (FID 0.248, R-Top1 0.512); the perturbation (Eq. 8) restores local diversity around the retrieved noise, mitigating the distribution shift caused by discrete selection. Subsequent refinement (WINRO) further strengthens alignment: for MDM-50step, the FID decreases to 0.115 and R-Top1 rises to 0.539; for MotionLCM, WINRO achieves

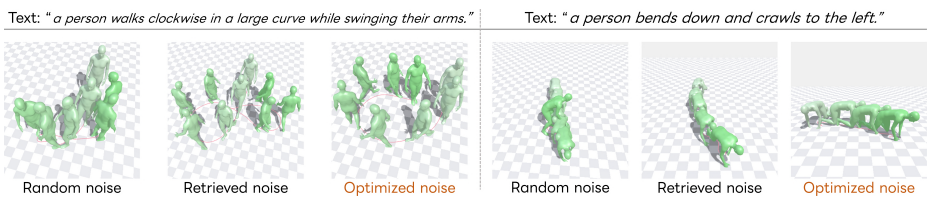


Fig. 7: Comparison of motions generated with random, retrieved, and refined initial noise for each text prompt, using MDM-50step with and without our method. The texts shown above each row are used for generation. We recommend viewing the supplementary video for better visualization of motion dynamics.

the best R-Top1 of 0.580 and MMDist of 2.609. Since the optimization objective already incorporates KL regularization (Eq. 11), we omit additional perturbation during refinement. LoRA-based refinement with retrieval and perturbation (WIN-RLoRA) provides an amortized alternative to full WINRO refinement, which is iterative and suited for offline or quality-critical use. Although WIN-RLoRA requires an additional training step and is therefore not training-free, it avoids iterative test-time optimization. It further improves over WINRO on MDM-50step and remains competitive on MotionLCM, while reducing AITS from 81.408 to 0.280 and from 6.870 to 0.043, respectively.

Note that the retrieval model (TMR [26]) used for dictionary construction and refinement is independent of the evaluation model used for the HumanML3D metrics. We further verify that improvements are not specific to the choice of retrieval backbone by repeating the experiment with an alternative encoder; details are provided in the supplementary material. The noise dictionary is constructed once offline per backbone: approximately 4 minutes for MDM-50step and 2 minutes for MotionLCM on a single A100 GPU, after which per-query inference requires only retrieval (negligible) and optional refinement. Additional analyses and ablations are provided in the supplementary material.

Qualitative Results. Fig. 7 compares motions generated with random noise, retrieved noise (WIN-FAR), and refined noise (WINRO) using MDM-50step.

In the first example, the prompt demands a clockwise walk. Random noise produces a generic walking motion with no clear trajectory, while the retrieved noise captures a partial arc but eventually turns in the wrong direction. In contrast, WINRO generates a smooth clockwise arc that closely matches the intended motion. In the second example, the prompt requires the person to bend down and crawl to the left. Random noise yields a crawling motion that ends in an unintended standing pose, and the retrieved noise produces only slight leftward motion. With WINRO refinement, the model is able to generate a consistent leftward crawl as instructed.

These qualitative results illustrate the complementary roles of retrieval and refinement: retrieval selects a noise seed with an inherent bias toward the desired behavior, while refinement further adjusts it to satisfy the prompt more faithfully. Additional examples are in the supplementary material.

Table 3: Results on the MTT benchmark [27].

Method	Input type		Per-crop semantic correctness				Realism	
	#tracks	#crops	R@1 ↑	R@3 ↑	M2T ↑	M2M ↑	FID ↓	Trans. ↓
Ground Truth	–	–	55.0	73.3	0.748	1.000	0.000	1.5
MDM-50step	Single	Single	12.2	24.6	0.571	0.561	0.543	2.2
w/ WINRO	Single	Single	**20.7**	**38.1**	**0.661**	**0.597**	**0.492**	**2.1**
w/ STMC [27]	Multi	Multi	33.6	55.3	0.687	0.656	0.480	1.7
w/ STMC + WIN-O	”	”	**62.1**	**79.1**	**0.807**	**0.702**	**0.452**	**1.4**

5.3 Compositional and Long-Duration Generation.

Practical motion generation often requires composing multiple actions within a single sequence: actions may be arranged in temporal order (*e.g.*, “walk forward, then sit down”) or performed simultaneously on different body parts (*e.g.*, “wave the right hand while walking”). The multi-track timeline (MTT) formulation [27] unifies these settings by assigning each text prompt to a temporal interval on parallel tracks, allowing overlapping and sequential actions to be specified in a single structured input. Generating faithful motions from such timelines is challenging because the model must maintain semantic alignment across all intervals, especially as the total duration grows. WINRO naturally fits this setting, since it can also be applied to multiple interval-level text constraints.

Method. We evaluate on the MTT benchmark [27] under two settings: (1) single-text compositional generation ($\leq$196 frames) using MDM-50step, where the full WINRO pipeline is applied, and (2) multi-track long-duration generation using STMC [27], where only the optimization stage is used because constructing a noise dictionary from variable-length null-prompt outputs is computationally impractical at extended durations. For the multi-track setting, the similarity loss is defined as the average text–motion similarity over timeline intervals. Each motion subsequence is encoded and compared with its corresponding text prompt, enabling the optimization to maintain semantic consistency across segments.

Results. As shown in Table 3, both configurations yield consistent improvements, with particularly large gains in long-duration generation (R@1: 33.6 $\rightarrow$ 62.1). These results suggest that winning noise tickets carry semantic priors that remain effective even in extended and compositional settings, where semantic drift is otherwise difficult to control. Fig. 8 shows qualitative examples: STMC alone suffers from semantic drift in later segments, whereas WIN-O maintains faithful alignment with each text prompt throughout the timeline.

5.4 Task Adaptability

As noted in Sec. 4, the retrieval and optimization stages are modular: the dictionary can be replaced with a domain-specific one or omitted, and the optimization

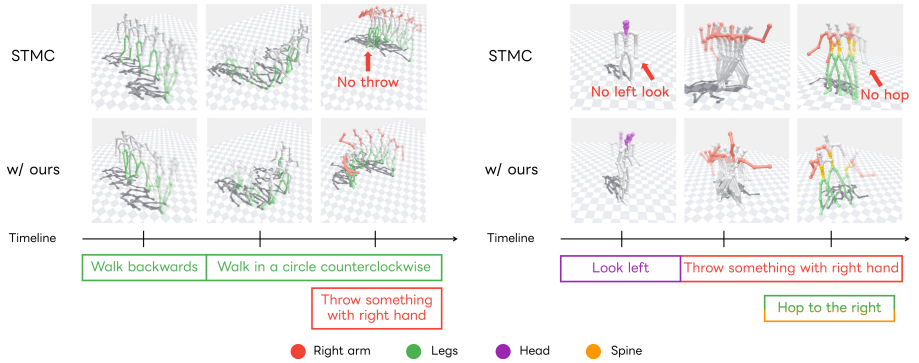


Fig. 8: Comparison of motions generated by STMC [27] with and without initial noise refinement on multi-track long-duration generation. STMC produces motions that are not faithful to the text prompt of each segment (*e.g.*, missing throw or hop actions), while refining the initial noise improves alignment.

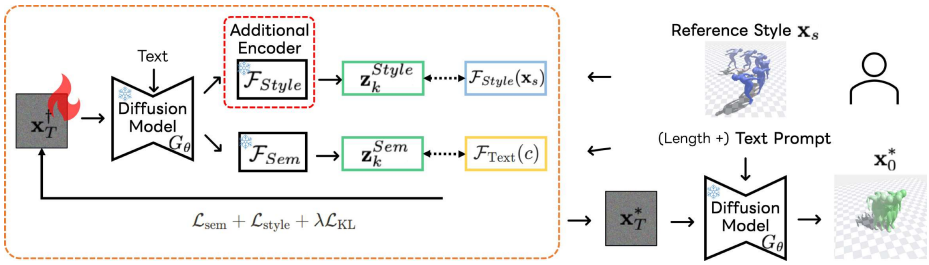


Fig. 9: Overview of WINRO for motion stylization. The generic noise dictionary is replaced with a style-specific one built from 100STYLE motions [23] via DDIM inversion. The optimization jointly minimizes semantic loss $\mathcal{L}_{\text{sim}}$, style loss $\mathcal{L}_{\text{style}}$, and KL regularization term $\lambda\mathcal{L}_{\text{KL}}$.

objective can incorporate additional loss terms. We demonstrate this flexibility on two tasks.

5.4.1 Motion Stylization. Motion stylization aims to generate motions that follow a given text prompt while exhibiting a particular movement style (*e.g.*, performing a "walking" motion in a "proud" or "depressed" manner), guided by a reference style motion.

Method. We extend WINRO by constructing a style-specific noise dictionary $\mathcal{D}_s$: we apply DDIM inversion to motions from the 100STYLE dataset [23], so that each entry inherently encodes stylistic attributes. Given a text prompt, we retrieve the best-matching entry from $\mathcal{D}_s$ using the same text-based retrieval as in Eq. 7, then refine the retrieved noise by jointly optimizing the semantic loss $\mathcal{L}_{sim}$ (Eq. 11) and a style loss $\mathcal{L}_{style} = 1 - \cos(\mathcal{F}_{\text{Style}}(\mathbf{x}_0^\dagger), \mathcal{F}_{\text{Style}}(\mathbf{x}_s))$, where

Table 4: Effect of WINRO on motion stylization using SMooDi as the backbone.

Method	FID ↓	Foot Skating ↓	MMDist ↓	R-Top3 ↑	Diversity →	SRA ↑
SMooDi [42]	1.609	0.124	4.477	0.571	9.235	72.418
w/ WIN-O	**0.551**	**0.104**	**3.550**	**0.717**	**9.071**	68.029
w/ WINRO	1.066	0.129	4.055	0.633	7.766	**75.818**

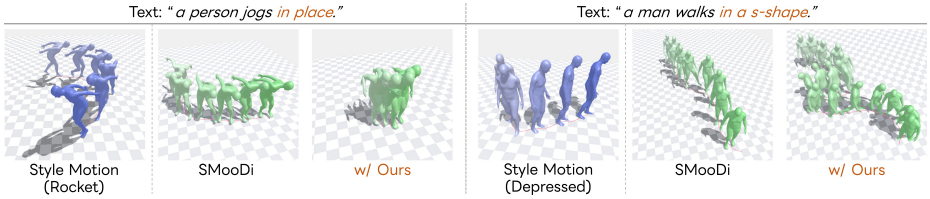


Fig. 10: Qualitative results from the stylization experiment. Reference style motion at the left and the text prompt at the top of each example are used for generation.

$\mathbf{x}_s$ is the reference style motion. We compare WIN-O (optimization only from random noise) and WINRO (retrieval from $\mathcal{D}_s$ + optimization). Fig. 9 illustrates the pipeline. Since WINRO is model-agnostic, it can be applied to any diffusion-based stylization backbone. We adopt SMooDi [42] for its public availability.

Results. Table 4 shows that both variants improve upon the SMooDi backbone without modifying its parameters. WIN-O, starting from random noise, optimizes freely toward text alignment and achieves the best FID (0.551 vs. 1.609), MMDist, and R@3. WINRO, starting from the style-specific dictionary, begins closer to the target style; accordingly, it attains the highest Style Recognition Accuracy (SRA) (75.82 vs. 72.42), confirming stronger style adherence, while text-alignment metrics remain competitive. The two variants thus offer complementary trade-offs: WIN-O prioritizes text fidelity, while WINRO prioritizes style preservation—both operating solely at the noise level.

Qualitative Results We qualitatively compare stylized motions in Fig. 10. In the first example, both SMooDi and our method reproduce the "Rocket" style, but SMoodi's output drifts semantically, producing a wandering motion instead. Our framework selects a more suitable initial noise, yielding a coherent "jogging in place" motion that respects both style and text. A similar trend is observed in the second example, where the baseline captures style but loses semantic context, while our method maintains alignment between content and style.

5.4.2 Controllable Motion Generation. Beyond text semantics, practical applications often require motions to satisfy explicit spatial constraints, such as reaching a specific location or avoiding obstacles. The optimization objective can incorporate such geometric or physical constraints from prior works [17,

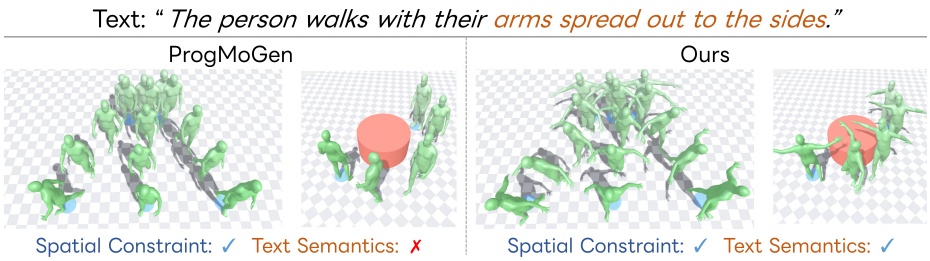


Fig. 11: Comparison of motions generated using the initial noise optimized by ProgMoGen [20] and the refined noise produced by the proposed WINRO method. WINRO produces motions that better satisfy both textual and spatial constraints.

20]. Concretely, we augment Eq. 11 with a constraint term $\beta\,\mathcal{L}_{\text{ctrl}}$ (*e.g.*, the waypoint/target-reaching objective of ProgMoGen [20]). As shown in Fig. 11, this combined objective finds an initial noise that satisfies both textual semantics and external spatial constraints, such as reaching target locations.

6 Limitations

Since WINRO does not modify the base model, it cannot address systematic failure modes of the diffusion backbone and is limited to guiding generation within its learned manifold. The iterative refinement stage incurs substantial inference overhead, as each optimization step requires a full diffusion generation pass; the LoRA-based refiner mitigates this for latency-sensitive applications, but requires an additional training step. Finally, the current motion-text retrieval model treats motions holistically, making it difficult to pinpoint fine-grained temporal properties in noise patterns. Investigating retrieval models with temporal localization is a promising direction for future work.

7 Conclusion

We present WINRO, a training-free and model-agnostic framework that exploits the semantic structure of the initial noise space to improve diffusion-based text-to-motion generation. By retrieving and refining winning noise tickets, WINRO consistently improves text–motion alignment across MDM and MotionLCM without retraining the base model. WINRO also improves temporal robustness on the MTT benchmark. Its modular design generalizes to motion stylization and spatial constraint satisfaction, demonstrating that principled noise selection is a versatile and complementary tool for controllable motion synthesis.

Acknowledgements

This study was carried out using the TSUBAME4.0 supercomputer at Institute of Science Tokyo.

References

1. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 3DV (2022)
2. Chen, C., Yang, L., Yang, X., Chen, L., He, G., Wang, C., Li, Y.: Find: Fine-tuning initial noise distribution with policy optimization for diffusion models. In: ACMMM (2024)
3. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR (2023)
4. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. In: ECCV (2024)
5. Eyring, L., Karthik, S., Dosovitskiy, A., Ruiz, N., Akata, Z.: Noise hypernetworks: Amortizing test-time compute in diffusion models. In: NeurIPS (2025)
6. Eyring, L., Karthik, S., Roth, K., Dosovitskiy, A., Akata, Z.: Reno: Enhancing one-step text-to-image models through reward-based noise optimization. In: NeurIPS (2024)
7. Frankle, J., Carbin, M.: The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: ICLR (2019)
8. Fujiwara, K., Tanaka, M., Yu, Q.: Chronologically accurate retrieval for temporal grounding of motion-language models. In: ECCV (2024)
9. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: ICCV (2021)
10. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR (2022)
11. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: CVPR (2024)
12. Holden, D., Kanoun, O., Perepichka, M., Popa, T.: Learned motion matching. TOG **39**(4), 53–1 (2020)
13. Hong, S., Kim, C., Yoon, S., Nam, J., Cha, S., Noh, J.y.: Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In: CVPR (2025)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
15. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. In: NeurIPS (2023)
16. Kalakonda, S.S., Maheshwari, S., Sarvadevabhatla, R.K.: Action-gpt: Leveraging large-scale language models for improved and generalized zero shot action generation. arXiv preprint arXiv:2211.15603 (2022)
17. Karunratanakul, K., Preechakul, K., Aksan, E., Beeler, T., Suwajanakorn, S., Tang, S.: Optimizing diffusion noise can serve as universal motion priors. In: CVPR (2024)
18. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: CVPR (2020)
19. Li, Z., Wang, S., Zhang, Z., Tang, H.: Remomask: Retrieval-augmented masked motion generation. arXiv preprint arXiv:2508.02605 (2025)
20. Liu, H., Zhan, X., Huang, S., Mu, T.J., Shan, Y.: Programmable motion generation for open-set motion control tasks. In: CVPR (2024)
21. Mao, J., Wang, X., Aizawa, K.: Guided image synthesis via initial image editing in diffusion model. In: ACMMM (2023)
22. Mao, J., Wang, X., Aizawa, K.: The lottery ticket hypothesis in denoising: Towards semantic-driven initialization. In: ECCV (2024)

23. Mason, I., Starke, S., Komura, T.: 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 (2022)
24. Ota, S., Yu, Q., Fujiwara, K., Ikehata, S., Sato, I.: Pino: Person-interaction noise optimization for long-duration and customizable motion generation of arbitrary-sized groups. In: ICCV (2025)
25. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: ICCV (2021)
26. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: ICCV (2023)
27. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Peng, X.B., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation. In: CVPRW (2024)
28. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big Data **4**(4), 236–252 (2016)
29. Shin, S., Kim, J., Halilaj, E., Black, M.J.: WHAM: Reconstructing world-grounded humans with accurate 3D motion. In: CVPR (2024)
30. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
31. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
32. Wang, R., Huang, H., Zhu, Y., Russakovsky, O., Wu, Y.: The silent prompt: Initial noise as implicit guidance for goal-driven image generation. arXiv preprint arXiv:2412.05101 (2024)
33. Xu, K., Zhang, L., Shi, J.: Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. arXiv preprint arXiv:2405.14828 (2024)
34. Xu, L., Song, Z., Wang, D., Su, J., Fang, Z., Ding, C., Gan, W., Yan, Y., Jin, X., Yang, X., Zeng, W., Wu, W.: Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In: ICCV (2023)
35. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: CVPR (2023)
36. Yu, Q., Tanaka, M., Fujiwara, K.: Exploring vision transformers for 3d human motion-language models with motion patches. In: CVPR (2024)
37. Yu, Q., Tanaka, M., Fujiwara, K.: Remogpt: Part-level retrieval-augmented motion-language models. In: AAAI (2025)
38. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: ECCV (2020)
39. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: CVPR (2023)
40. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE TPAMI (2024)
41. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: ICCV (2023)
42. Zhong, L., Xie, Y., Jampani, V., Sun, D., Jiang, H.: Smoodi: Stylized motion diffusion model. In: ECCV (2024)