MixiMotion: One-Step Text-to-Motion Generation via Asymmetric Set Distillation

Hung Dinh[1 [0009−0003−6278−1092]], Binh Mai[1 [0009−0009−9957−8702]], Tran Quoc Bao Le[1 [0009−0001−9024−5858]], Lam Nguyen[1 [0009−0005−3550−9864]], and Cong Tran[1 [0000−0001−9467−4978]⋆]

Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
{dinhhung15082004, binhmai2205, k100iltqbao, hailam04.work}@gmail.com, congtt@ptit.edu.vn

Abstract. Diffusion and flow-matching models synthesize high-quality 3D human motion from text, yet iterative samplers require tens of network evaluations. Collapsing the sampler to one step is not a scheduling problem: one prompt admits many valid motions, so regression onto a single teacher sample averages modes, and latent residuals still decode into skating or root drift. We present MixiMotion, a strict one-step generator distilled from an offline teacher cache. A frozen multi-step model is run once per prompt to store several motion candidates; the student maps independent noises to one-step motions and is trained by asymmetric bidirectional set matching plus decoded kinematic constraints. The teacher is never queried during optimization. On ViMoGen, MixiMotion scores 0.880 versus 0.886 for the 50-step teacher under a shared renderer, at one network evaluation and 9.3 ms versus 830 ms on an RTX 5090. Project page: https://paper.dinhmanhhung.net/

Keywords: Text-to-motion · Knowledge distillation · One-step generation · Human motion synthesis · Semantic alignment

1 Introduction

Text-to-motion generation maps natural language to temporally coherent 3D human motion. Diffusion and flow models such as MDM [24], MLD [2], and HY-Motion-1.0 [26] have improved motion quality and semantic fidelity. Their iterative samplers, however, require many network function evaluations (NFEs), which limits interactive authoring, game-engine, and web deployment.

Reducing the solver to one evaluation is not merely a scheduling problem. For one text prompt, several motions can be equally valid, so pointwise regression to one target encourages conditional averaging. Errors that are small in a normalized motion representation may still produce visible foot skating, weak steps, or root drift after decoding. A practical one-step student must therefore cover several valid teacher motions and remain plausible after decoding.

⋆ Corresponding author: Cong Tran (congtt@ptit.edu.vn).

In this paper, MixiMotion caches multiple multi-step teacher outputs per prompt and trains a one-step student with asymmetric bidirectional set matching and decoded body-space constraints (Fig. 1). The teacher is frozen HY-Motion-1.0-Lite [26], the compact flow-matching generator in the HY-Motion family; we run it once per prompt to fill the cache and never call it during student training. Reusing the Lite backbone is an implementation choice, not an architectural contribution. We train on the full ViMoGen prompt bank and test on ViMoGen. Under a shared renderer, MixiMotion obtains 0.8802 versus 0.8856 for 50-NFE HY-Motion-1.0 full. The paired interval includes zero. Human ratings still favor the iterative reference, and MBench generalizability is substantially lower, so the result is a quality–latency trade-off inside one model family, not a general one-step advance.

Our contributions are:

- An asymmetric set-matching objective for strict one-step text-to-motion that covers multiple cached teacher motions without online teacher sampling.
- Decoded kinematic supervision for contact, gait, and root behavior that latent distance alone does not see.
- A ViMoGen train/test comparison under a shared renderer, with contextual cross-family numbers, human ratings, and a remaining generalizability gap.

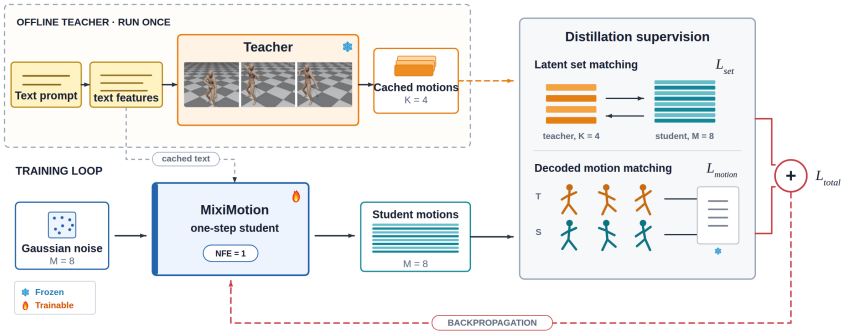

Fig. 1. MixiMotion overview. A frozen multi-step teacher produces K cached targets per prompt once. The one-step student generates M candidates from independent noise samples and is optimized by latent and decoded-motion set matching; only the student receives gradients.

2 Related Work

Text-to-motion generation. T2M-GPT [28] models discrete motion tokens autoregressively. MDM [24], MotionDiffuse [29], and MLD [2] introduced diffusion in motion or latent spaces. MotionGPT [14] treats motion as a foreign language for unified motion–language modeling. ReMoDiffuse [30] retrieves motion exemplars to guide denoising. StableMoFusion [12] studies robust diffusion backbones

and training-free step reduction. Motion Mamba [31] targets long-sequence efficiency with state-space blocks. HY-Motion-1.0 [26] scales flow-matching transformers. These iterative or reduced-step models provide strong references but remain expensive at large NFE budgets.

Accelerated generative models. Knowledge distillation transfers teacher behavior to a student [9]. Progressive distillation [22] and consistency models [23] reduce iterative sampling by teaching a student to replace several teacher steps or to agree along a trajectory. Rectified flows learn straighter transport paths that admit few Euler steps [18]. Distribution matching distillation [27] and score-identity distillation [32] train one-step generators by matching diffused teacher and student distributions rather than paired trajectories. In text-to-audio, SwiftAudio [20] uses caption-only variational score distillation and temporal smoothness regularization for one-step generation without paired audio supervision. MixiMotion instead learns from an offline cache of teacher-generated motions through set matching. For motion, MotionLCM [3] supports one- to four-step latent-consistency sampling, and MotionPCM [15] develops phased consistency training, whereas MotionHiFlow [16] uses a reduced hierarchical flow-matching solver. Unlike MotionLCM, MixiMotion does not learn a consistency function along a diffusion trajectory: it maps noise to motion in one evaluation and is supervised by an offline multi-candidate teacher set. MoFlow [6] applies one-step IMLE distillation to human trajectory forecasting; it motivates sample-based matching but is a different task and is not a text-to-motion baseline. We report the actual NFE of every evaluated configuration instead of conflating strict one-step and reduced-step regimes.

Set matching and ranking. IMLE [17] replaces direct likelihood optimization with nearest-neighbor matching between observed and generated samples. We use an asymmetric, bidirectional variant to cover the cached teacher set while discouraging unsupported student samples. The logistic relative-margin loss of Direct Preference Optimization [21] is a pairwise ranking form against a frozen reference. Our last stage reuses only that form: pair labels come from a hand-specified kinematic score on cached teacher motions, not from human preferences or model likelihoods.

3 Method

MixiMotion distills a strict one-step generator from an offline multi-candidate teacher bank. The recipe does not assume a particular diffusion or flow backbone: any frozen multi-step text-to-motion model that emits a motion sequence can serve as teacher, and any one-step map in the same output space can serve as student.

3.1 Offline Teacher Bank and One-Step Student

Let G_T be a frozen multi-step teacher and G_θ a one-step student. For prompt p and seeds $\xi_1, \ldots, \xi_K$, the teacher is evaluated only while building the cache:

$$\mathcal{Y}_T(p) = \{y_j = G_T(p; \xi_j)\}_{j=1}^{K}. \tag{1}$$

Each $y_j \in \mathbb{R}^{T \times D}$ is a motion sequence in the teacher's normalized representation, of valid length $\ell \le T$. The bank also stores the conditioning required by G_T, written $c(p)$. For independent $z_i \sim \mathcal{N}(0, I)$,

$$\mathcal{X}_\theta(p) = \{\hat{y}_i = G_\theta(z_i, c(p))\}_{i=1}^{M}. \tag{2}$$

G_θ predicts a clean motion in one network evaluation. A frozen body decoder F_{dec} maps a sequence to 3D joints, root translation, and root orientation for kinematic supervision. Gradients update only θ. The integers K, M, T, and D are instance choices (Sec. 4).

3.2 Asymmetric Set Distillation

Text-to-motion is multi-modal, so we match sets rather than a single pair. Let ρ_δ be a robust elementwise penalty (Smooth-L1 in our run [13]). For a frame subset $S \subseteq \{1, \ldots, \ell\}$,

$$d_\delta^S(a, b) = \frac{1}{|S|D} \sum_{t \in S} \sum_{d=1}^{D} \rho_\delta(a_{t,d} - b_{t,d}). \tag{3}$$

Matching from teacher to student covers every cached target. The reverse direction penalizes unsupported samples. For a cost matrix $C \in \mathbb{R}^{K \times M}$,

$$\mathcal{B}_{\boldsymbol{\alpha}}(C) = \frac{\alpha_{\mathrm{t2s}}}{K} \sum_{j=1}^{K} \min_i C_{ji} + \frac{\alpha_{\mathrm{s2t}}}{M} \sum_{i=1}^{M} \min_j C_{ji} \tag{4}$$

is an asymmetric Chamfer-style reduction [5], not a bijection.

Let $S_{\mathrm{all}} = \{1, \ldots, \ell\}$ and $S_{\mathrm{end}} = \{\ell\}$. Trajectory costs subsample P frames along the clip:

$$S_{\mathrm{traj}} = \left\{ 1 + \mathrm{round}\left(\frac{q(\ell - 1)}{P - 1} \right) \right\}_{q=0}^{P-1}. \tag{5}$$

The pairwise costs in the teacher's representation are then

$$C_{ji}^{\mathrm{lat}} = d_\delta^{S_{\mathrm{all}}}(y_j, \hat{y}_i), \quad C_{ji}^{\mathrm{end}} = d_\delta^{S_{\mathrm{end}}}(y_j, \hat{y}_i), \quad C_{ji}^{\mathrm{traj}} = d_\delta^{S_{\mathrm{traj}}}(y_j, \hat{y}_i). \tag{6}$$

Let $F_{\mathrm{dec}}(y)$ be the decoded body trajectory. A decoded map Φ_r is any differentiable feature of that trajectory (body, foot, and root coordinates; contact; stride; lift). Write $\Omega_r(\ell)$ for the index set of its valid scalar entries at length ℓ. Contact labels are computed from the teacher so the student cannot define its

own contacts. The pair cost of map r averages a Huber penalty with threshold δ_{joint} over $\Omega_r(\ell)$:

$$C_{ji}^{(r)} = \frac{1}{|\Omega_r(\ell)|} \sum_{u \in \Omega_r(\ell)} \rho_{\delta_{\text{joint}}}\big([\Phi_r(y_j)]_u - [\Phi_r(\hat{y}_i)]_u\big). \tag{7}$$

The family $\mathcal{R}$ is skeleton- and decoder-dependent; any differentiable maps in the decoded space may be substituted. Nonnegative weights $\{\lambda_r\}_{r\in\mathcal{R}}$ form $C^{\text{body}} = \sum_{r\in\mathcal{R}} \lambda_r C^{(r)}$. The base objective is

$$\mathcal{L}_{\text{set}} = \mathcal{B}_{\boldsymbol{\alpha}}(C^{\text{lat}}) + \lambda_{\text{end}}\mathcal{B}_{\boldsymbol{\alpha}}(C^{\text{end}}) + \lambda_{\text{traj}}\mathcal{B}_{\boldsymbol{\alpha}}(C^{\text{traj}}) + \mathcal{B}_{\boldsymbol{\alpha}}(C^{\text{body}}). \tag{8}$$

All coefficients are nonnegative. Which maps are active, and whether decoded or latent matching dominates, is a weighting choice rather than a change of objective.

3.3 Optional Staged Refinement

The same bank can be reused under a sequence of weightings (Fig. 2). Later stages are not assumed to improve every aggregate metric.

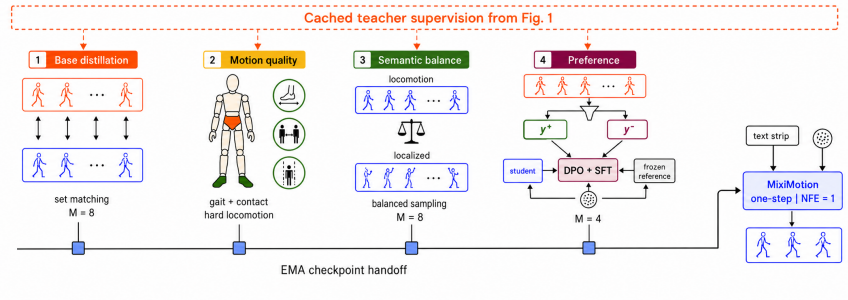

Fig. 2. One staged schedule used in our run. Arrows encode checkpoint order and do not imply monotonic metric improvement. Other teachers may omit stages or change group weights.

Decoded-quality emphasis. Write $F_{\text{dec}}(y) = (x, \tau, \psi)$ for joints, root translation, and root yaw. Let j_{L} and j_{R} be the left and right foot joints of the chosen skeleton, $f \in \{\text{L}, \text{R}\}$, and $x_{t,j}^{xz}$ the ground-plane coordinates of joint j at frame t. Planar stride is the left–right foot separation

$$\Phi_{\text{stride}}(x)_t = \|x_{t,j_{\text{L}}}^{xz} - x_{t,j_{\text{R}}}^{xz}\|_2. \tag{9}$$

The ground-relative height $h_{t,f}$ is the vertical clearance of foot f above the lowest foot point in the clip (a clip-wise ground proxy):

$$h_{t,f} = \max\left(0,\, x_{t,j_f}^{y} - \min_{t',f'} x_{t',j_{f'}}^{y}\right). \tag{10}$$

Soft contact follows the height-threshold convention used for foot-contact and skating statistics in text-to-motion evaluation [10,7,24]:

$$c_{t,f} = \sigma\left(\frac{h_{\mathrm{c}} - h_{t,f}}{\tau_{\mathrm{c}}}\right), \tag{11}$$

so $c_{t,f}$ approaches 1 when the foot is near the ground. Alternating contact (exactly one foot down) and either-contact (at least one foot down) are

$$a_t^{\mathrm{c}} = c_{t,\mathrm{L}}(1 - c_{t,\mathrm{R}}) + (1 - c_{t,\mathrm{L}})c_{t,\mathrm{R}}, \qquad e_t^{\mathrm{c}} = 1 - (1 - c_{t,\mathrm{L}})(1 - c_{t,\mathrm{R}}). \tag{12}$$

Foot skating is planar travel while in contact [10,7]:

$$\Phi_{\mathrm{skate}}(x)_{t,f} = c_{t,f}\, \|x^{xz}_{t+1,j_f} - x^{xz}_{t,j_f}\|_2. \tag{13}$$

Here σ is the logistic sigmoid and $(h_{\mathrm{c}}, \tau_{\mathrm{c}})$ are skeleton-scale thresholds. Root maps use planar translation increments and wrapped yaw increments. A stage that emphasizes decoded matching uses

$$\mathcal{L}_{\mathrm{dec}} = \mathcal{B}_{\boldsymbol{\alpha}}(C^{\mathrm{body}}) + \lambda_{\mathrm{lat}}\mathcal{B}_{\boldsymbol{\alpha}}(C^{\mathrm{lat}}) + \lambda_{\mathrm{end}}\mathcal{B}_{\boldsymbol{\alpha}}(C^{\mathrm{end}}) + \lambda_{\mathrm{traj}}\mathcal{B}_{\boldsymbol{\alpha}}(C^{\mathrm{traj}}). \tag{14}$$

Prompt groups (locomotion versus stationary) may be oversampled. For prompts that should not travel, student-only root-stability terms

$$\mathcal{L}_{\mathrm{rd}}(p) = \frac{1}{2M\ell}\sum_{i=1}^{M}\sum_{t=1}^{\ell}\sum_{c\in\{x,z\}} \rho_{\delta_{\mathrm{joint}}}(\hat{\tau}^c_{i,t} - \hat{\tau}^c_{i,1}), \tag{15}$$

$$\mathcal{L}_{\mathrm{rv}}(p) = \frac{1}{2M(\ell-1)}\sum_{i=1}^{M}\sum_{t=1}^{\ell-1}\sum_{c\in\{x,z\}} \rho_{\delta_{\mathrm{joint}}}(\hat{\tau}^c_{i,t+1} - \hat{\tau}^c_{i,t}) \tag{16}$$

are added on a subset $\mathcal{P}_{\mathrm{stat}}$:

$$\mathcal{L}_{\mathrm{sem}}(p) = \mathcal{L}_{\mathrm{dec}}(p) + \mathbf{1}[p \in \mathcal{P}_{\mathrm{stat}}]\big(\lambda_{\mathrm{rd}}\mathcal{L}_{\mathrm{rd}}(p) + \lambda_{\mathrm{rv}}\mathcal{L}_{\mathrm{rv}}(p)\big). \tag{17}$$

Heuristic ranking. If several teacher motions remain, a scalar $q_{g(p)}$ on decoded kinematics (lift, contact, skating, stride, root travel) can label a pair

$$y^+ = \arg\max_{y_j \in \mathcal{Y}_T(p)} q_{g(p)}(y_j), \qquad y^- = \arg\min_{y_j \in \mathcal{Y}_T(p)} q_{g(p)}(y_j), \tag{18}$$

kept when the score gap exceeds a margin η. A frozen reference copy θ_{ref} and the student share the same noises. With $s_\theta(y) = -\min_i d^{S_{\mathrm{all}}}_{\delta_{\mathrm{lat}}}(\hat{y}_i, y)$,

$$\begin{aligned}\mathcal{L}_{\mathrm{rank}} &= -\log\sigma\big(\beta\left[s_\theta(y^+) - s_\theta(y^-) - s_{\mathrm{ref}}(y^+) + s_{\mathrm{ref}}(y^-)\right]\big),\\ \mathcal{L}_{\mathrm{pref}} &= \mathcal{L}_{\mathrm{rank}} + \lambda_{\mathrm{sft}}\min_i d^{S_{\mathrm{all}}}_{\delta_{\mathrm{lat}}}(\hat{y}_i, y^+).\end{aligned} \tag{19}$$

This reuses a relative-margin form [21]. It is not human preference learning: q is an instance heuristic on the teacher cache, and s_θ is a motion distance.

Weights, optimizer, and stage lengths for the reported checkpoint are given in Sec. 4.

4 Experiments

4.1 Setup

Data. The prompt bank is the full set of 50,000 ViMoGen text annotations [25]; no HY-Motion example text or additional prompt source is included. HY-Motion-1.0-Lite generates $K = 4$ deterministic 120-frame targets per prompt. Every training stage and ranking pair uses this bank. Testing uses ViMoGen (1,969 semantic prompts and MBench).

Instantiation. The reported student uses HY-Motion-1.0-Lite as G_T [26] and reuses that model's multimodal DiT interface as G_θ (one-step clean prediction; 460.0M generator parameters). This is an implementation choice, not part of the method. We set $K = 4$, $T = 120$, $D = 201$, $P = 4$, and $M = 8$ ($M = 4$ in the ranking stage). For this backbone, $c(p)$ is the cached text context, global text vector, and length mask. Motion and text-context inputs (201 and 4096) and a 768-D global text vector are projected to 1024-D; six double-stream and twelve single-stream MM-DiT blocks follow. Contact uses teacher ankle/toe joints with speed 0.02 m/s and heights $(0.12, 0.12, 0.05, 0.05)$ m. Gait maps use foot joints $(10, 11)$, $h_c = 0.05$, and $\tau_c = 0.02$. Stationary root weights are $\lambda_{rd} = 0.80$ and $\lambda_{rv} = 0.30$. The ranking score $q_{g(p)}$ is a linear combination of lift, contact, skating, stride, and root travel; pairs with gap below $\eta = 0.03$ are dropped. Training uses AdamW [19], batch size 4, bfloat16, EMA 0.999, $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, $\delta_{lat} = 0.35$, $\delta_{joint} = 0.04$, $\beta = 8$, and $\lambda_{sft} = 0.35$. Tables 1 and 2 list decoded weights and stage lengths for this run.

Table 1. Decoded-feature weights in this run. Entries give position/velocity/acceleration ("p/v/a"), contact velocity ("c"), yaw velocity ("y"), and stride/lift/contact-pattern ("s/l/c"). A dash denotes zero weight.

Phase	Body p/v/a	Foot p/v/c	Key p/v	Root v/y	Gait s/l/c	Arm p/v
Base	.20/.12/.04	.30/.25/.32	.18/.12	–	–	–
Gait/body	.16/.10/.03	.24/.18/.18	.16/.10	.10/.04	.30/.55/.40	.14/.08
Hard loco.	.14/.10/.04	.22/.20/.14	.14/.10	.12/.04	.38/.85/.65	.10/.06
Semantic	.14/.10/.04	.20/.18/.12	.14/.10	.10/.04	.30/.60/.45	.10/.06

Table 2. Training schedule for the reported checkpoint. "Lat." is λ_{lat} in (14). Later stages are not selected by the test metric.

Phase	Epochs	Learning rate	M	Key setting
Base distillation	18	8×10^{-5}	8	End/traj. .10/.12; $\boldsymbol{\alpha} = (1, .20)$
Gait/body matching	10	3×10^{-5}	8	Lat. .12; end/traj. .04/.06; $\boldsymbol{\alpha} = (1, .15)$
Hard locomotion	6	2×10^{-5}	8	Lat. .08; end/traj. .02/.04; 6×
Semantic balance	4	1×10^{-5}	8	Lat. .08; end/traj. .02/.04; 4×
Heuristic ranking	3	2×10^{-6}	4	Margin + SFT; shared noise

Protocols. We train on ViMoGen and test on ViMoGen. Semantic scores are the yes-rate of fine-grained questions on video, judged by Qwen3-VL-30B-A3B-Instruct-FP8 [1] with all 120 frames, temperature zero, and a fixed yes/no schema. Every configuration has 1,969 successfully judged videos and 6,418 answered questions. We evaluate strict one-step MixiMotion and MotionLCM-1; few-step MotionLCM-2/4 [3] and MotionHiFlow [16]; and multi-step HY-Motion-1.0 full, MLD [2], MDM [24], MotionDiffuse [29], and UMF [11]. UMF is a number-free multi-person framework and is included only as additional context.

MixiMotion and HY-Motion full share the same 720×720 renderer, camera, 30 FPS encoding, prompt indices, and sequence length; this is the only renderer-controlled quality comparison. Public methods retain their native renderers and frame rates, so their semantic scores are contextual and are not a model-only ranking. For efficiency, all four profiled systems use one RTX 5090, BF16, batch size 1, 20 warm-ups, and 100 synchronized timed runs. Generation latency covers the model-native solver. End-to-end latency also includes each method's text encoder and decoder, whose endpoints differ across families.

MBench [25] contains 450 unique prompts: 150 trajectory-quality prompts shared by five joint-based metrics, 100 prompts shared by body penetration and pose quality, 100 motion-condition-consistency prompts, and 100 motion-generalizability prompts. Every evaluated configuration generates all 450 motions. For trajectory quality, outputs are converted to a common 20 FPS, 22-joint representation and scored with the official evaluator. Body penetration and pose quality use the benchmark's official SMPLify, BVH self-collision, and NRDF pipeline on their prescribed 100 prompts. For the latter two semantic dimensions, all systems use the same 512×512 neutral-skeleton renderer and the same all-frame Qwen judge. These Qwen values substitute for the official Gemini judge and are not directly comparable with the official MBench leaderboard.

For human evaluation, 25 raters assess 300 ViMoGen prompts per system under randomized, blinded presentation. They score semantic alignment, naturalness, and overall motion quality on five-point Likert scales. We report mean±standard error (SE) and Krippendorff's α following the reliability formulation of Hayes and Krippendorff [8].

4.2 ViMoGen Semantic Results

Same-family semantic alignment. The intended ViMoGen comparison is MixiMotion versus 50-NFE HY-Motion-1.0 full under a shared renderer (Table 3). The absolute gap is 0.0055. MixiMotion ties the reference on locomotion at the reported precision. A paired percentile bootstrap [4] with 10,000 video-level resamples gives MixiMotion−HY-full = −0.0055 with 95% CI $[-0.0121, 0.0011]$. The interval includes zero, so the judge does not resolve a significant difference; this is not an equivalence test. Cross-family scores, including MotionHiFlow at 0.8764 and MotionLCM-1 at 0.8415, use native renderers. Those rows are deployment context, not evidence that MixiMotion is the best one-step model in general. Under the same mixed protocol, MixiMotion−MotionHiFlow is −0.0092

Table 3. ViMoGen semantic scores under the all-frame native-video judge. Panel (a) ranks one-step and few-step methods only. Panel (b) is a separate multi-step reference. MixiMotion versus HY-Motion-1.0 full is renderer-controlled; other rows keep native renderers. Bold is best *within each panel*.

(a) One-step and few-step

Method	NFE↓	Daily↑	Fit.↑	Game↑	Loco.↑	Social↑	Sports↑	Overall↑
MixiMotion (ours)	1	**0.9009**	0.7487	**0.9205**	**0.9757**	0.9332	**0.8430**	**0.8802**
MotionLCM-1	1	0.8065	0.7213	0.8977	0.9621	0.9211	0.7807	0.8415
MotionHiFlow	14	0.8617	**0.7717**	0.9187	0.9689	**0.9452**	0.8274	0.8764
MotionLCM-2	2	0.7963	0.7405	0.8932	0.9621	0.9168	0.7838	0.8429
MotionLCM-4	4	0.8019	0.7257	0.8968	0.9670	0.9189	0.7838	0.8425

(b) Multi-step reference

Method	NFE↓	Daily↑	Fit.↑	Game↑	Loco.↑	Social↑	Sports↑	Overall↑
HY-Motion-1.0 full	50	**0.9075**	**0.7613**	**0.9215**	**0.9757**	**0.9419**	**0.8451**	**0.8856**
MLD	50	0.8112	0.7079	0.8959	0.9582	0.9310	0.7796	0.8398
UMF	32	0.8103	0.7287	0.8932	0.9456	0.9124	0.7630	0.8364
MDM	50	0.8196	0.7235	0.8648	0.9504	0.9266	0.7516	0.8331
MotionDiffuse	1000	0.7822	0.6887	0.8548	0.9261	0.8992	0.7422	0.8087

(95% CI $[-0.0186, 0.0002]$) and MixiMotion−MotionLCM-1 is -0.0441 (95% CI $[-0.0553, -0.0332]$).

4.3 ViMoGen Test (MBench)

Among one- and few-step systems, MixiMotion has the lowest body-penetration error (1.0752) and the best NRDF pose quality (2.7641). Condition consistency is 0.38, versus 0.46 for the multi-step HY-Motion full reference. Motion generalizability is 0.24 versus 0.40 for HY-Motion full and 0.30 for MotionHiFlow: a one-step coverage failure, not a residual gap. Trajectory metrics among one-/few-step methods are mixed: MixiMotion is best on jitter and foot sliding, while MotionHiFlow is best on ground penetration and foot floating. MBench therefore does not support uniform physical quality, nor that set matching recovers the teacher distribution.

4.4 Efficiency

MotionLCM-1 is the smallest and fastest profiled system (26.5M parameters, 3.04 ms generation, 13.94 ms end-to-end). MixiMotion is a 460.0M one-step student (44.1% of HY-Motion-full parameters, 20.73 GiB). The 89.2× generation and 6.75× end-to-end speedups are only versus 50-NFE HY-Motion full under matched-family timing. They are not an efficiency win against compact public one-step models. Cross-family end-to-end numbers remain deployment context because text encoders and decoded endpoints differ.

Table 4. Official MBench physical metrics. Panels (a)–(b) rank one-step and few-step methods only. Panels (c)–(d) are a separate multi-step reference. Trajectory uses 150 prompts; body/pose use 100 prompts and the official SMPLify/NRDF/BVH pipeline. Semantic scores use a common Qwen judge and are not comparable with the official Gemini leaderboard. Bold is best *within each panel*. Dynamic degree is descriptive.

(a) One-step and few-step: trajectory

Method	NFE	Jitter ↓	Ground penetr. ↓	Foot float. ↓	Foot slide ↓	Dynamic ↔
MixiMotion (ours)	1	**.0198**	.0008	.3186	**.0047**	.0410
MotionLCM-1	1	.0219	.0009	.1743	.0077	.0427
MotionLCM-2	2	.0225	.0013	.1939	.0075	.0440
MotionLCM-4	4	.0214	.0010	.1658	.0074	.0434
MotionHiFlow	14	.0223	**.0002**	**.1464**	.0064	.0456

(b) One-step and few-step: pose and semantic

Method	NFE↓	Body penetration ↓	Pose quality ↓	Condition consistency ↑	Motion generalizability ↑
MixiMotion (ours)	1	**1.0752**	**2.7641**	**0.38**	0.24
MotionLCM-1	1	1.8141	2.8199	0.30	0.28
MotionLCM-2	2	1.8108	2.8227	0.31	0.25
MotionLCM-4	4	1.7222	2.8480	0.33	0.27
MotionHiFlow	14	1.4394	2.9289	0.29	**0.30**

(c) Multi-step reference: trajectory

Method	NFE	Jitter ↓	Ground penetr. ↓	Foot float. ↓	Foot slide ↓	Dynamic ↔
HY-Motion-1.0 full	50	.0333	**.0005**	.2475	.0068	.0506
MLD	50	.0175	.0007	.1309	.0059	.0425
UMF	32	.0192	.0033	.4476	.0144	.0404
MDM	50	.0120	.0009	**.0808**	**.0036**	.0339
MotionDiffuse	1000	**.0111**	.0018	.0827	.0041	.0295

(d) Multi-step reference: pose and semantic

Method	NFE↓	Body penetration ↓	Pose quality ↓	Condition consistency ↑	Motion generalizability ↑
HY-Motion-1.0 full	50	1.3496	2.9578	**0.46**	**0.40**
MLD	50	1.5805	2.8040	0.24	0.26
UMF	32	1.7310	2.7927	0.27	0.26
MDM	50	1.8690	3.0184	0.33	0.18
MotionDiffuse	1000	**1.3413**	**2.7477**	0.25	0.25

Table 5. Efficiency on one RTX 5090 in BF16 (batch size 1, 120 frames). Parameter counts cover generation checkpoints. Cross-family end-to-end endpoints are model-native; ↓ indicates lower is better.

(a) Model complexity and memory

Method	Generator params (M)↓	Actual NFE↓	Peak VRAM (GiB)↓
HY-Motion-1.0 full	1042.9	50	23.23
MixiMotion (ours)	460.0	1	20.73
MotionHiFlow	33.1	14	**0.43**
MotionLCM-1	**26.5**	1	1.45

(b) Latency

Method	Gen. mean±std↓	Gen. P95↓	E2E mean±std↓	E2E P95↓
HY-Motion-1.0 full	829.58±3.19	834.68	968.25±3.69	974.94
MixiMotion (ours)	9.30±0.39	9.94	143.55±1.29	145.67
MotionHiFlow	77.32±1.33	79.73	83.80±1.17	85.82
MotionLCM-1	**3.04±0.24**	**3.35**	**13.94±0.47**	**14.85**

Table 6. Blinded human ratings on a five-point scale (300 prompts per system; mean±SE; ↑ is better). Bold denotes the best overall; underlining denotes the best one-/few-step result. Krippendorff's $\alpha = 0.68$.

Method	Alignment↑	Naturalness↑	Overall quality↑
HY-Motion-1.0 full	$\mathbf{4.48 \pm 0.04}$	$\mathbf{4.52 \pm 0.03}$	$\mathbf{4.50 \pm 0.04}$
MixiMotion (ours)	$\underline{4.32 \pm 0.05}$	$\underline{4.35 \pm 0.04}$	$\underline{4.33 \pm 0.05}$
MotionHiFlow	4.18 ± 0.06	4.22 ± 0.05	4.20 ± 0.06
MotionLCM-1	3.95 ± 0.07	4.01 ± 0.06	3.98 ± 0.07

4.5 Human and Qualitative Evaluation

Human means still rank HY-Motion full first. MixiMotion is next among the four rated systems, but the 0.16–0.17 gaps to the iterative reference are descriptive; no pairwise significance claim is made from these aggregate SEs. Selected qualitative frames in Fig. 3 are successful examples, not a random sample, and do not include compact one-step baselines.

4.6 Limitations

The student reuses the teacher backbone, so this is not an architecture result. All systems are evaluated at 120 frames. The human study covers a finite prompt and rater sample. Qwen-scored MBench semantic dimensions cannot be compared with the official Gemini leaderboard. SMPLify body and pose scores depend on inverse-kinematics fitting. MBench generalizability is substantially below the multi-step reference. Broader claims would need independently sourced prompts, a common cross-family renderer, and component ablations on the reported checkpoint.

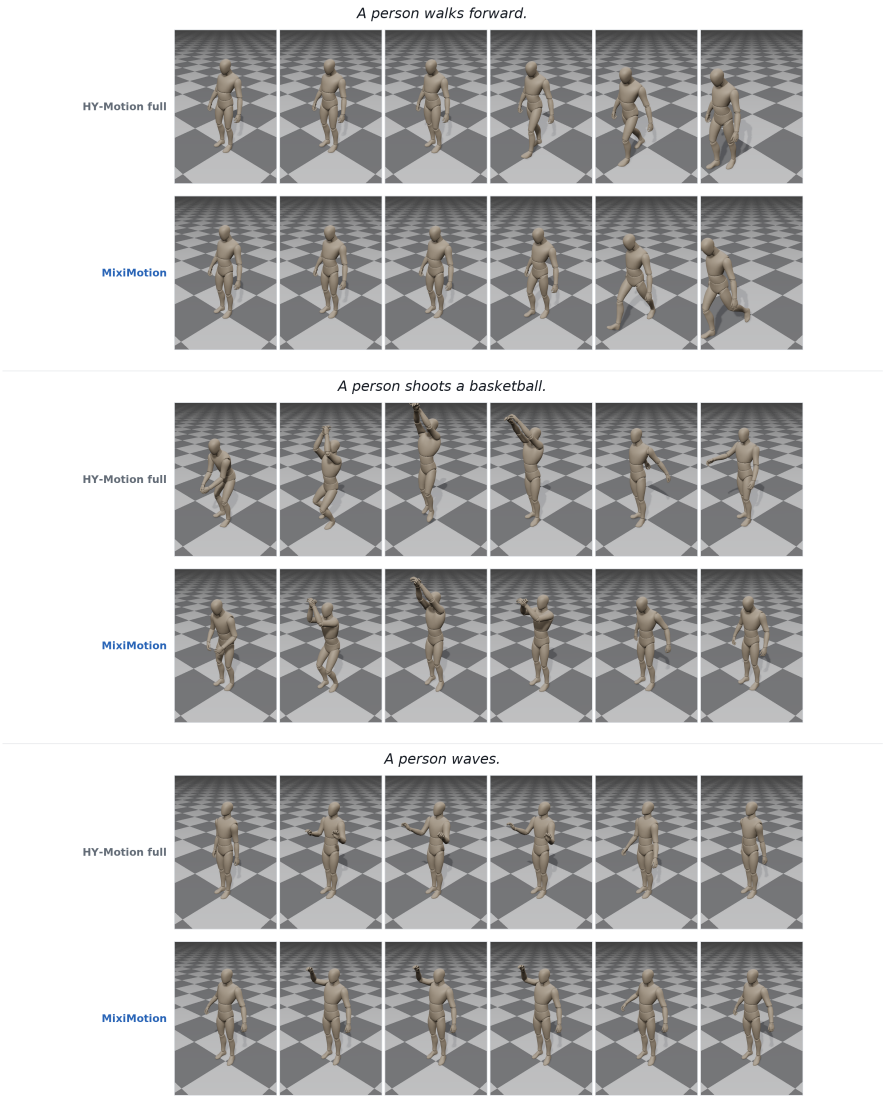

Fig. 3. Qualitative comparison on three selected successful instructions. For each prompt, six temporally ordered frames from the 120-frame sequence are shown for HY-Motion-1.0 full (upper row) and MixiMotion (lower row). Both systems use the same renderer, camera, and lighting. The examples illustrate temporal progression but are not a random sample.

5 Conclusion

MixiMotion is a strict one-step text-to-motion student trained by offline asymmetric set matching and decoded kinematic supervision. It reuses the teacher backbone and does not introduce a new generator architecture. Training uses the full ViMoGen prompt bank; testing uses ViMoGen. Under a shared renderer it scores 0.8802 versus 0.8856 for 50-NFE HY-Motion-1.0 full; the paired interval includes zero. Matched-family generation latency falls by 89.2×. MotionLCM-1 remains smaller and faster. Human ratings still favor the iterative reference. MBench generalizability (0.24 vs. 0.40) shows that one-step set matching does not recover the teacher distribution. The result is a same-family quality–latency trade-off, not general one-step superiority.

Acknowledgments. The authors thank the maintainers of HY-Motion-1.0, ViMoGen, and the open-source text-to-motion ecosystem.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, S., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
3. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: MotionLCM: Real-time controllable motion generation via latent consistency model. In: European Conference on Computer Vision (2024)
4. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall/CRC (1993)
5. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 605–613 (2017)
6. Fu, Y., Yan, Q., Wang, L., Li, K., Liao, R.: MoFlow: One-step flow matching for human trajectory forecasting via implicit maximum likelihood estimation based distillation. arXiv preprint arXiv:2503.09950 (2025)
7. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: 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)
8. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Communication Methods and Measures **1**(1), 77–89 (2007)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
10. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Transactions on Graphics **35**(4), 138:1–138:11 (2016)
11. Huang, G., Celiktutan, O.: Unified number-free text-to-motion generation via flow matching. arXiv preprint arXiv:2603.27040 (2026)

12. Huang, Y., Yang, H., Luo, C., Wang, Y., Xu, S., Zhang, Z., Zhang, M., Peng, J.: StableMoFusion: Towards robust and efficient diffusion-based motion generation framework. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 224–232 (2024)
13. Huber, P.J.: Robust estimation of a location parameter. The Annals of Mathematical Statistics **35**(1), 73–101 (1964)
14. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: MotionGPT: Human motion as a foreign language. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
15. Jiang, L., Wei, Y., Ni, H.: MotionPCM: Real-time motion synthesis with phased consistency model. arXiv preprint arXiv:2501.19083 (2025)
16. Li, H., Lin, X., Zeng, L.A., Kang, Y., Li, S., Hu, J.F.: MotionHiFlow: Text-to-motion via hierarchical flow matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
17. Li, K., Malik, J.: Implicit maximum likelihood estimation. arXiv preprint arXiv:1809.09087 (2018)
18. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (2023)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
20. Mai, B., Le, T.Q.B., Dinh, H., Tran, C.: SwiftAudio: Data-efficient caption-only distillation for one-step text-to-audio diffusion-based generation. arXiv preprint arXiv:2606.31259 (2026), `https://arxiv.org/abs/2606.31259`
21. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
22. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022)
23. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: Proceedings of the 40th International Conference on Machine Learning. pp. 32211–32252 (2023)
24. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: International Conference on Learning Representations (2023)
25. ViMoGen-228K: Large-scale video-to-motion prompt and motion dataset. Hugging Face dataset repository (2026), `https://huggingface.co/datasets/wruisi/ViMoGen-228K`
26. Wen, Y., Shuai, Q., Kang, D., Li, J., Wen, C., Qian, Y., Jiao, N., Chen, C., Chen, W., Wang, Y., et al.: HY-Motion 1.0: Scaling flow matching models for text-to-motion generation. arXiv preprint arXiv:2512.23464 (2025)
27. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
28. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
29. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: MotionDiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)

30. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: ReMoDiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 364–373 (2023)
31. Zhang, Z., Liu, A., Reid, I., Hartley, R., Zhuang, B., Tang, H.: Motion mamba: Efficient and long sequence motion generation. In: European Conference on Computer Vision (2024)
32. Zhou, M., Zheng, H., Wang, Z., Yin, M., Huang, H.: Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In: Proceedings of the 41st International Conference on Machine Learning (2024)