TRACE: TYPED REFERENCE-AWARE CONVERSATIONAL EDITING FOR HUMANOID MOTION DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-conditioned humanoid motion generation turns a textual request into a sequence of robot poses. A short request leaves timing, support, and posture partly unspecified, yet a useful motion must realize these choices coherently and retain them when the user asks for a local change. We present TRACE, a framework for motion generation and reference-aware editing on the Unitree G1 humanoid robot. Our central idea is to express motion structure and requested revisions through typed kinematic quantities that can be grounded in a motion reference. During training, forward kinematics extracts support, root, and hand events from paired motions to supervise a language model. At inference, the language model predicts duration and event intervals from text; their numerical encodings are aligned to motion patches and combined with the original caption in a diffusion Transformer. A gravity-referenced representation directly predicts height and tilt, while trajectory and contact objectives constrain accumulated motion errors. For revision, a typed request specifies a body–time scope and a goal relative to the current motion. A separate diffusion editor uses source features and revised events to realize that goal, with reference projection preserving protected coordinates throughout sampling. Experiments show strong text–motion alignment and human ratings. Matched ablations support structured event conditioning and improved kinematic consistency, while shared-scope comparisons demonstrate more reliable completion of successive edits than the adapted learned editors.

1 INTRODUCTION

Language-conditioned humanoid motion generation maps a natural-language instruction to a sequence of root poses and joint configurations. It lets users specify behavior such as walking, dancing, or jumping without authoring each pose. The resulting trajectory can serve as an animation or a reference for a downstream robot controller. In interactive use, generation also provides a starting point for refinement: after seeing a motion, a user may ask for a higher jump or a different hand position while expecting the accepted parts of the sequence to remain recognizable.

Existing text-to-motion models learn mappings from captions to continuous trajectories or discrete motion codes (Tevet et al., 2023; Zhang et al., 2023a; Jiang et al., 2023; Guo et al., 2024). Their text features capture broad action semantics, while support changes and local timing remain largely implicit in the generated trajectory. Structured approaches expose body-part or temporal conditions (Zhang et al., 2023b; Yazdian et al., 2025), and language-model planning decomposes instructions into more detailed descriptions (Ouyang et al., 2026). Such descriptions can still specify a plausible realization different from the paired training motion: a caption alone rarely determines exactly when a foot supports the body or how high a hand rises. Geometric objectives and simulation-based methods address motion consistency (Tevet et al., 2023; Yuan et al., 2023; Tevet et al., 2025), while source-conditioned editors learn changes from language (Athanasiou et al., 2024; Li et al., 2025; Lin et al., 2026). For an interactive humanoid system, these capabilities must connect semantic intent to measurable motion structure and to the parts of a reference that a revision should retain.

The *key challenge* is to turn underspecified language into a coherent motion reference whose intended properties can be revised selectively. Figure 1 illustrates this requirement. An initial instruction produces a jumping jack; the follow-up "Jump higher, rest unchanged" refers to that particular motion. The edited example crouches deeper and jumps higher, requiring coordinated changes in

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.

preparation and take-off while retaining the overall action. A useful intermediate description should therefore expose quantities such as support and height over time, and a revision should be defined relative to the motion already accepted by the user. Our insight is to ground these quantities in the paired demonstration during training and in the current reference during editing.

We introduce *Typed Reference-Aware Conversational Editing* (TRACE) for the Unitree G1 humanoid robot. *Typed* refers to explicit categories for support, root, and hand events and for requested changes. *Reference-aware* means that the current motion defines both the baseline for a relative change and the coordinates to preserve. For generation, we extract event targets from training motions through forward kinematics and supervise an LLM to predict duration and event intervals from text. The predicted Plan is converted into numerical conditions aligned with motion patches. A diffusion Transformer combines these conditions with the original caption, retaining language information beyond the compact event vocabulary. A gravity-referenced state directly represents height and tilt, while integrated-trajectory and support-selective contact objectives constrain motion consistency.

For editing, a language request is parsed into a typed change and compiled into a body–time scope and a source-relative goal. A separately trained diffusion editor receives the current motion, its revised events, and the compiled goal. At every sampling step, reference projection restores protected coordinates, allowing the selected region to change while retaining the rest of the encoded motion. Each output supplies the reference for the next request. Generation and editing thus share a motion representation and a kinematic vocabulary, with separate models for their respective conditioning tasks.

Evaluation on BONES-SEED (Bones Studio, n.d.) combines generation comparisons, matched component ablations, and controlled editing tests. The results support the semantic value of motion-grounded event conditions and the kinematic value of the representation and consistency objectives. The whole-body example in Figure 1 provides qualitative evidence of refinement; quantitative editing uses hand-height requests with measurable relative goals and protected regions. Under shared scopes and projection, TRACE completes successive revisions more reliably than the adapted learned editors. Our contributions are:

1. **Motion-grounded typed-event conditioning.** We derive confidence-weighted support, root, and hand event targets from training trajectories and learn a text-to-Plan predictor. Patch-aligned numerical events complement raw-caption features in continuous G1 motion generation.
2. **Gravity-referenced motion diffusion.** We directly represent absolute height and gravity-relative tilt, and combine integrated-trajectory objectives with support-selective forward-kinematics penalties. Matched-budget controls evaluate the contribution of this representation and its objective bundles to kinematic consistency.
3. **Typed, source-relative motion editing.** We compile supported editing requests into body–time scopes and source-relative goals, and combine them with source-motion features and revised events in a separately trained editor. Reference projection preserves protected state coordinates. We evaluate request-to-scope compilation separately from sustained request completion under shared scopes and projection.

2 Related Work

Motion-grounded planning and conditioning. Text-to-motion models generate discrete codes (Zhang et al., 2023a; Jiang et al., 2023; Guo et al., 2024) or continuous trajectories (Tevet et al., 2023). Structured representations add finer control: MotionScript converts motion into detailed language (Yazdian et al., 2025), and FineMoGen models body-part and temporal composition (Zhang et al., 2023b). Motion-R1 synthesizes decomposed reasoning plans from descriptions and refines generation with RL (Ouyang et al., 2026). For supervised motion generation, the intermediate conditions also need to agree with the specific demonstration paired with the caption. TRACE addresses this alignment through confidence-weighted support, root, and hand events extracted from that training motion. A text-conditioned planner predicts these events at inference, and their numerical sequence supplements raw-caption features. Matched controls evaluate the event content and its temporal alignment.

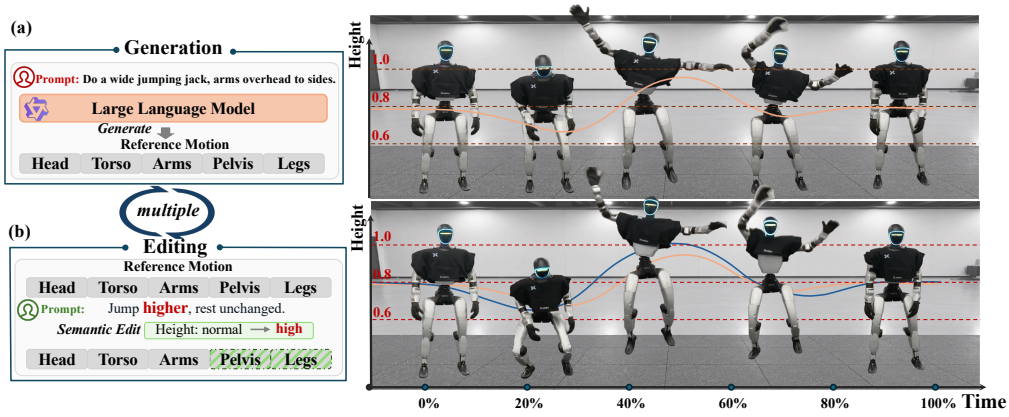


Figure 1: **From text generation to reference-aware refinement.** (a) The request "Do a wide jumping jack, arms overhead to sides" produces an initial motion reference. (b) "Jump higher, rest unchanged" specifies a relative height change; the schematic highlights the pelvis and legs involved in the revision. The robot sequence shows a deeper preparatory crouch and a higher jump. Orange and blue height profiles compare the initial and edited motions over normalized time.

Kinematic consistency and humanoid generation. MDM incorporates geometric and foot-contact objectives (Tevet et al., 2023). PhysDiff adds simulator-based motion projection during denoising (Yuan et al., 2023), while CLoSD closes the loop between a diffusion planner and a tracking controller (Tevet et al., 2025). Language-conditioned humanoid systems connect generation to robot motion or control (Jiang et al., 2024; Shao et al., 2025; Li et al., 2026; Jian et al., 2026); imitation and transfer methods address downstream tracking (Fu et al., 2024; He et al., 2024; 2025). These methods address motion plausibility and downstream execution at different stages. TRACE targets upstream reference quality through gravity-referenced height and tilt, integrated trajectory objectives, and support-selective FK contact losses.

Typed source-relative and reference-preserving editing. MDM demonstrates temporal and body-part inpainting (Tevet et al., 2023), and PriorMDM composes diffusion priors for joint and trajectory control (Shafir et al., 2024). FineMoGen also uses LLMs for fine-grained editing (Zhang et al., 2023b). MotionFix introduces source–instruction–target pairs and the source-conditioned TMED editor (Athanasiou et al., 2024); SimMotionEdit adds motion-similarity supervision (Li et al., 2025), and CIME studies spatial preservation and non-uniform temporal mapping (Lin et al., 2026). Together, these approaches establish source conditioning and spatial or temporal preservation as important parts of motion editing. TRACE studies their connection to explicit relative requests: a typed relative change is compiled into a kinematic body–time scope and coupled to source-relative goals, revised source events, and reference projection.

3 Method

3.1 Overview

Typed event plans structure generation; typed edit requests specify how an accepted motion should change. Reference-conditioned denoising realizes the change within its compiled scope, while projection preserves protected coordinates. TRACE uses separate generator and editor weights; Figure 2 depicts the generation pathway.

For an instruction u, a frozen text encoder gives $e = f(u)$, while a supervised LLM jointly predicts duration and typed event intervals. Deterministic rasterization converts its structured output into the plan:

$$P = (N, C) = \text{Rasterize}(\text{LLM}_\phi(u)), \quad C \in [0,1]^{L\times 16}, \quad L = \lceil N/4 \rceil. \tag{1}$$

The LLM predicts support, root, and hand events and is frozen during diffusion training and inference. We denote its rasterized events at length N by $h(u, N)$. A conditional diffusion Transformer generates continuous G1 states from (e, P).

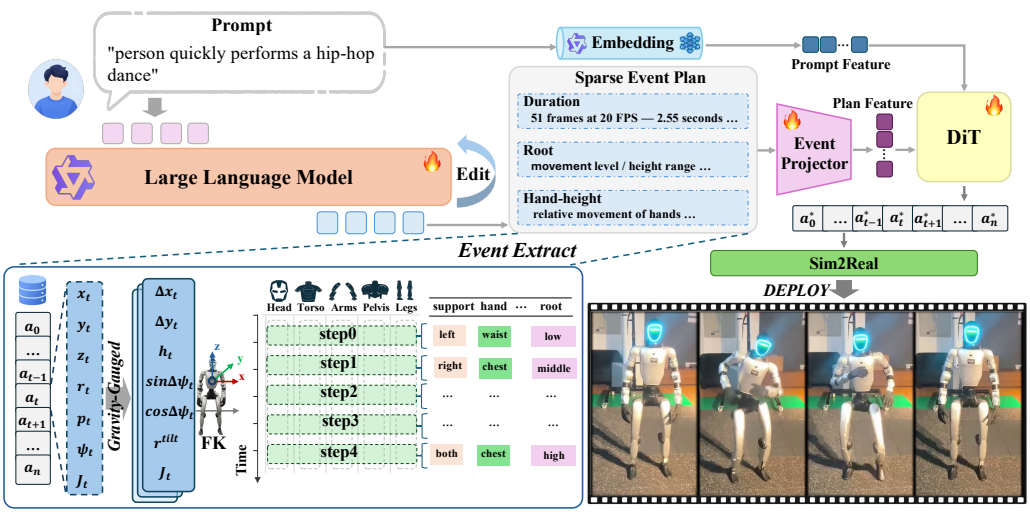


Figure 2: **TRACE with a motion-grounded sparse-event plan.** The upper path maps a text instruction to continuous G1 motion. A frozen text encoder supplies prompt features, while an LLM predicts duration and typed support, root, and hand events. The event projector maps their numerical, time-aligned encoding to plan features that condition the diffusion Transformer (DiT) alongside the prompt. The lower-left path constructs training supervision: gravity-gauged states and forward kinematics (FK) provide measurable signals for event extraction, linked to the plan by the dashed path. Test-time plans are predicted from text. Snowflake and flame symbols denote frozen and trainable modules in their respective training stages; the trained LLM is frozen for diffusion training. Generated references can be passed to a downstream controller for robot execution, illustrated at lower right. The edit loop denotes follow-up requests; Section 3.4 describes the separate reference-conditioned editor.

3.2 Motion-Grounded Sparse Planning and Conditioning

We construct a compact event vocabulary from measurable properties of the paired training motion. These motion-derived targets provide realization-aligned supervision for the language-to-plan predictor. At inference, predicted events supply explicit support, root, and hand structure, while the original caption retains semantics outside this vocabulary.

Motion-derived supervision. For each clean training motion, deterministic FK and root reconstruction produce sole heights/speeds, root vertical motion and height, and palm heights relative to the pelvis. Fixed thresholds convert these signals into typed events. Each four-frame patch carries two independent support probabilities and four categorical distributions:

$$c_\ell = [c_\ell^{L\text{foot}}, c_\ell^{R\text{foot}}, c_\ell^{\text{root-v}}, c_\ell^{\text{root-h}}, c_\ell^{L\text{hand}}, c_\ell^{R\text{hand}}], \qquad c_\ell \in [0,1]^{16}. \tag{2}$$

The groups define semantic types, not a programming-language type system. Categorical entries sum to one within each valid group. *Sparse* refers to the compact vocabulary; every valid motion patch receives a condition.

Plan prediction and training conditions. Threshold ambiguity produces a confidence W, not a forced negative label. Motion-derived pseudo-targets supervise the LLM through serialized typed plans using only training motions; Y denotes their numerical encoding. To reduce dependence on perfect annotations at deployment, diffusion training mixes target and predicted conditions while retaining raw text. Uncertain targets fall back to predicted events:

$$C^{\text{mix}} = \begin{cases} W \odot Y + (1-W) \odot h(u,N), & \text{with probability } 1/2, \\ h(u,N), & \text{otherwise.} \end{cases} \tag{3}$$

The choice applies jointly to all event groups for a sample. Training uses the recorded length; inference uses the caption-predicted length and predicted events only. Appendix B.1 specifies the extraction rules and an actual training example. These are kinematic pseudo-labels, not human-verified semantic annotations or a unique interpretation of an underspecified caption.

Time-aligned conditioning. The raw caption is encoded once. A learned numerical event projection and a separate raw-text residual branch use patch progress features to align conditions to the denoiser. Both supplement the raw-text backbone. At diffusion timestep k, the residual condition is scaled by $\alpha(k) = \min(1, k/[0.25(K-1)])$, attenuating it near clean motion. Matched controls retain the same branch capacity, so removing plan means zeroing its event inputs while preserving all trainable parameters.

3.3 Gravity-Gauged and Contact-Consistent Diffusion

The plan specifies motion structure, but accurate local poses alone do not ensure a coherent trajectory. Small errors in integrated root motion can accumulate, and plausible poses can still move a supporting foot. We therefore represent height and tilt relative to gravity and supervise integrated geometry and foot contact explicitly. This separates quantities that require an absolute reference from those that describe displacement.

Gravity-gauged states and kinematic consistency. We predict absolute height and gravity-relative tilt directly, integrating only planar displacement and yaw. The per-frame state is

$$x_t = [\Delta p_t^x, \Delta p_t^y, h_t, \sin \Delta\psi_t, \cos \Delta\psi_t, \mathbf{r}_t^{\text{tilt}}, j_t], \tag{4}$$

Here $\mathbf{r}_t^{\text{tilt}}$ concatenates the first two columns of $H_t^\top R_t$, where R_t is the body-to-world root rotation and H_t is its horizontal heading rotation; j_t contains 29 joint angles. The six tilt coordinates are orthonormalized and composed with integrated yaw to recover the root rotation, up to the canonical initial heading (Appendix C.1). This 40-dimensional gravity gauge is a representation, not a sensor input. Plan support probabilities condition the denoiser but are not additional predicted state coordinates.

Grouped training objective. The generator predicts clean normalized states under a cosine noise schedule. We organize its objective into three weighted groups:

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{motion}} + \mathcal{L}_{\text{geometry}} + \mathcal{L}_{\text{contact}}. \tag{5}$$

Motion reconstruction supervises states, velocities, and accelerations. *Geometry consistency* constrains reconstructed position, cumulative yaw, body-up direction, and the yaw-pair norm. *Contact consistency* uses differentiable FK to penalize support-foot sliding, drift, and height error, plus penetration on all valid frames. Clean training motion selects support masks; predicted plan probabilities are conditions, not substitutes for these masks. Each group includes its component weights, detailed in Appendix C.2.

3.4 Automatic Masking and Reference-Aware Motion Editing

Editing requires both locating the requested change and preserving the remaining motion. We combine automatic body–time masking, source-relative goals, and reference projection to address these complementary requirements.

Automatic scope localization. At turn t, $X_t = (M_t, u_t)$ contains the source and request; N_t is the source length. A supervised LLM, frozen at inference, predicts $A_t = \delta_t = \pi_\phi(u_t, N_t, s(M_t))$: an attribute, effector, signed change, and interval. The source-only summary $s(M_t)$ contains FK hand heights relative to the pelvis. A deterministic compiler constructs $(S_t, D_t) = g(A_t, u_t, N_t)$. For the body–time mask S_t, it expands the selected effector e_t to its kinematic chain $\mathcal{K}(e_t)$ and resolves interval $[a_t, b_t)$ against N_t:

$$S_t(f, j) = \mathbf{1}[j \in \mathcal{K}(e_t)]\, \mathbf{1}[a_t \le f < b_t]. \tag{6}$$

For hand-height edits, $\mathcal{K}$ contains seven arm-joint coordinates; its complement remains protected, including that arm outside the interval. Explicit time expressions are compiled from language; invalid requests leave the mask closed. No target motion or evaluation mask is supplied. The grammar is detailed in Appendix D.2.

Source-relative motion revision. For controlled hand-height evaluation, $D_t \in \mathbb{R}^{N_t \times 2}$ encodes signed wrist-height offsets with a sine-squared temporal envelope. The editor conditions on source-motion features, the scope, these offsets, and revised source events. The event sequence is extracted deterministically from the current source motion, with the requested hand-height groups updated according to D_t. Original and synthetic edit pairs supervise local motion fidelity, relative goals, and event consistency. This wrist-offset formulation quantifies hand edits, not the qualitative jump refinement in Figure 1. Appendix D.2 defines request compilation, and Appendix D.3 details the editor and its training.

Reference-preserving transitions. The conditional editor defines a stochastic state transition:

$$M_{t+1} \sim \mathcal{T}_\eta(\cdot \mid X_t, A_t) = p_\eta(\cdot \mid f(u_t), M_t, S_t, D_t). \tag{7}$$

Revised source events are deterministic functions of the source motion and compiled offsets; this dependence is implicit in the compact conditioning notation above. The next state is $X_{t+1} = (M_{t+1}, u_{t+1})$, with an external utterance; the Markov formulation assumes self-contained requests. After each DDIM step $\widetilde{x}_{k-1}$, the binary mask S preserves protected normalized coordinates by reference projection:

$$x_{k-1} = S \odot \widetilde{x}_{k-1} + (1 - S) \odot q_{k-1}(\mathrm{Enc}(M_t)). \tag{8}$$

The reference is noised to the current sampling level; the final step copies protected encoded state coordinates exactly. In the hand-height protocol, the root and nonselected joint chains remain fixed. For an edit that changes the root, preserving local joint coordinates would not by itself imply unchanged world-space positions of the corresponding body parts. Mask localization and editing under shared scopes are evaluated separately.

4 Motion Generation Experiments

After the setup (Section 4.1), we analyze generation quality (Section 4.2) and ablate Plan conditioning and kinematic design (Section 4.3).

4.1 Experimental Setup

Data. We use the processed BONES-SEED test split of 13,302 text–motion pairs. Each UID contributes one complete generated motion, including reassembly of packed segments where required. Processing, split integrity, and reassembly are detailed in Appendix E.1.

Baselines. We distinguish two groups. *Human-motion methods adapted to G1* comprise MDM (Tevet et al., 2023), MoMask (Guo et al., 2024), and MotionGPT3 (Zhu et al., 2026), retrained on the processed BONES-SEED/G1 split with heading-local 38D outputs. *Native G1 methods* comprise OMG (Huang et al., 2026), ECHO (Jia et al., 2026), TextOP (Xie et al., 2026), and TeXedo (Cao et al., 2026), using official checkpoints. TRACE is trained on the processed split and uses LLM-predicted Plans, never target-motion conditions. All methods are evaluated on the same test set; training and inference configurations are documented in Appendix E.2.

Metrics. All rows use the same frozen G1 text–motion evaluator for 32-candidate R-Precision, MM-Dist, FID, and Diversity. Diversity is a distributional diagnostic, not a ranked objective. Physical errors are assessed only in matched component controls. Table captions identify the uncertainty measure; Appendix E.4 gives definitions, repetitions, and aggregation.

Implementation. TRACE generates 40D G1 states at 20 FPS with four-frame patches. The diffusion Transformer has eight layers, width 448, and eight heads. Frozen Qwen3-Embedding-4B supplies caption features; a Qwen3-4B-Instruct adapter jointly predicts duration and all event groups. The main-table generator continues a 32k-step raw-text model for 8k steps with the 16-event branch and frozen LLM predictions. Test inference uses predicted conditions, EMA, DDIM-50, and CFG 2.5. Appendix E.2 records the full training and sampling configuration and batch sizes.

Table 1: BONES-SEED comparison: automatic metrics on test motions (R@k in percent) and a user study on 100 test prompts (scores 1–3). Cells show means above and uncertainty below where reported: evaluator-resampling 95% CI for R-Precision and SD across 16 participant means for the study. Automatic metrics and user-study scores follow the protocols defined in Appendix E.4.

	Automatic test-set metrics						User study		
Method	R@1↑	R@2↑	R@3↑	MM-Dist↓	FID↓	Diversity	Quality↑	Text match↑	Overall↑
TextOP (Xie et al., 2026)	10.54 ±0.07	19.02 ±0.09	25.18 ±0.11	1.2950	0.3985	1.2437	1.331 ±0.199	1.281 ±0.152	1.306 ±0.154
TeXedo (Cao et al., 2026)	10.78 ±0.10	17.80 ±0.09	22.60 ±0.07	1.3150	0.4060	1.2505	1.275 ±0.191	1.106 ±0.106	1.191 ±0.125
ECHO (Jia et al., 2026)	24.08 ±0.12	36.90 ±0.13	44.05 ±0.13	1.2008	0.2577	1.3288	1.469 ±0.218	1.469 ±0.196	1.469 ±0.195
OMG (Huang et al., 2026)	25.48 ±0.10	38.51 ±0.13	45.19 ±0.11	1.2158	0.2921	1.3060	1.356 ±0.287	1.300 ±0.193	1.328 ±0.200
MoMask (Guo et al., 2024)	31.57 ±0.08	41.88 ±0.10	46.80 ±0.07	1.1801	0.3057	1.2716	1.463 ±0.258	1.250 ±0.216	1.356 ±0.188
MDM (Tevet et al., 2023)	58.84 ±0.09	73.51 ±0.11	79.31 ±0.06	0.9716	0.0897	1.3769	1.313 ±0.225	1.313 ±0.193	1.313 ±0.182
MotionGPT3 (Zhu et al., 2026)	68.46 ±0.12	81.29 ±0.09	85.81 ±0.08	0.8886	0.0628	1.3775	1.538 ±0.236	1.300 ±0.146	1.419 ±0.155
TRACE (Ours)	**73.84** ±0.13	**86.24** ±0.08	**90.31** ±0.07	**0.8165**	**0.0434**	1.3766	**2.338** ±0.410	**2.444** ±0.407	**2.391** ±0.401

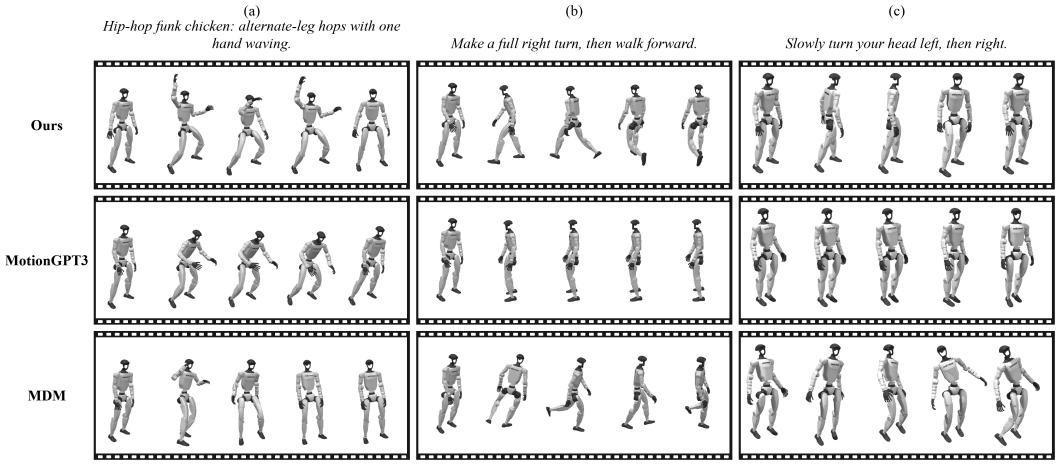


Figure 3: **Qualitative comparison of Ours, MotionGPT3, and MDM.** The third column illustrates generation from a free-form text prompt.

4.2 Generation Performance Analysis

TRACE achieves the best retrieval, MM-Dist, and FID results in Table 1. Retrieval measures caption compatibility through ranking, MM-Dist measures paired embedding distance, and FID measures distributional agreement with held-out motions. Their consistent improvement indicates strong semantic alignment and distributional fidelity of the complete system, while the user study provides complementary evidence of perceptual quality and text alignment.

Figure 3 compares TRACE with MotionGPT3 and MDM. The selected test examples qualitatively illustrate closer semantic alignment through clearer dance gestures and turn-then-walk ordering. Since Table 1 evaluates the complete TRACE system, we do not attribute these gains to any single component here. The contribution of motion-grounded event conditioning is examined separately through the matched controls in Section **??**.

4.3 Ablation Study

Motion-Grounded Plan Ablation. We evaluate the effect of Plan conditioning and its supervision source with matched controls (Figure 4). Aligned Plan uses the main-table model; Text Only retains the same raw-caption features, branch capacity, duration predictor, initialization, objectives, and continuation budget while zeroing the event inputs. A second comparison replaces motion-derived auxiliary targets with caption-derived events during training and retains the same predicted Plan

at test time. The displayed retrieval and distributional improvements support typed, time-aligned conditioning and supervision grounded in the paired motion. The former exposes when support and height events occur; the latter keeps the auxiliary targets consistent with the trajectory being reconstructed, whereas a caption-derived Plan can describe another valid realization. These generation metrics provide evidence for the conditioning design; exact event adherence requires a separate measurement. Test conditions are predicted from text throughout. Appendix B.2 provides the Global Events control, absolute values, and full protocols.

Kinematic Design Ablation. We evaluate the effect of the representation and objective bundles using Heading-local, Gravity-Gauged, Geometry, and Contact models trained for the same budget on the same data and event conditions (Figure 5). Gravity-Gauged reduces contact-height error, consistent with predicting absolute height and tilt directly. Geometry further improves trajectory-related diagnostics but increases penetration; adding Contact reduces penetration and yields lower errors than Heading-local across all reported measures. This pattern indicates that integrated-trajectory supervision and foot–ground constraints address complementary errors. Each plotted metric is normalized by its Heading-local mean. Contact diagnostics use reference support intervals, and penetration uses all valid frames; absolute values and training details appear in Appendix C.3.

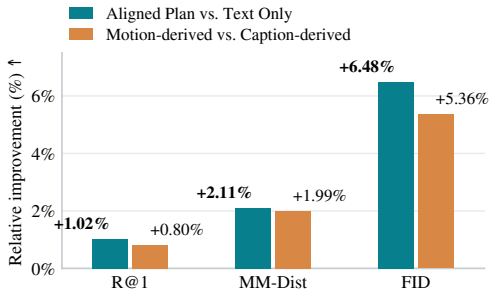


Figure 4: **Motion-grounded Plan ablation.** Teal compares Aligned Plan with Text Only; orange compares motion-derived with caption-derived training supervision under a shared predicted test Plan. Relative improvement is $100(v-b)/b$ for retrieval and $100(b-v)/b$ for MM-Dist/FID. Positive is better, not percentage-point change. Protocol details are in Appendix B.2.

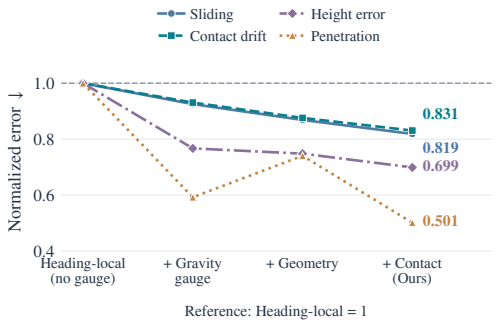


Figure 5: **Reference-based kinematic evaluation.** Each error is normalized by its Heading-local mean (baseline 1; lower is better). The first transition changes the state representation; Geometry and Contact subsequently add objective bundles. Contact-based measures use reference support intervals, while penetration is evaluated over all valid frames.

5 Motion Editing Experiments

We separately test automatic localization and multi-turn execution under shared masks.

5.1 Experimental Setup

Data and tasks. We use hand-height editing as a controlled test of source-relative request execution. A requested height change defines a measurable FK goal while allowing multiple valid joint trajectories, so evaluation can assess goal achievement and preservation without requiring a unique target motion. Within this protocol, the signed offset follows a fixed sine-squared profile over the requested interval. This profile is the benchmark's temporal convention, not a uniquely inferred interpretation of the request. We use 199 held-out BONES-SEED/G1 source motions with human-reviewed synthetic English editing requests. Each initial request specifies a hand, a signed 5 cm height change, and an explicit time interval. We first evaluate the predicted body–time mask without generating motion. For sequential editing, each source yields three scripted four-turn patterns: same-arm edits, alternating arms, and reversals, giving 597 chains.

Models and controls. For motion editing, we compare the complete TRACE editor with local G1 adaptations of TMED (Athanasiou et al., 2024) and SimMotionEdit (Li et al., 2025) under a shared scope-compilation and reference-projection protocol.

Metrics. Mask localization uses body–time IoU against private reference masks. All-turn success requires every request through a depth to meet change and preservation criteria relative to its actual preceding output (Appendix E.4).

5.2 Scope-Construction Diagnostics

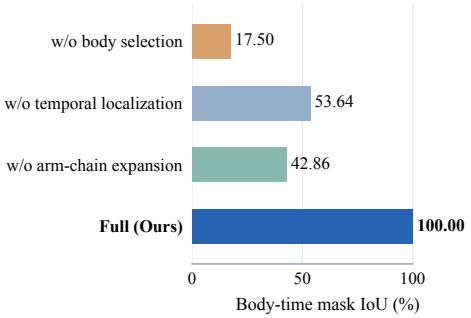


Figure 6: **Scope diagnostics.** Body–time IoU on 199 reviewed requests; parser outputs are fixed while compiler components are removed.

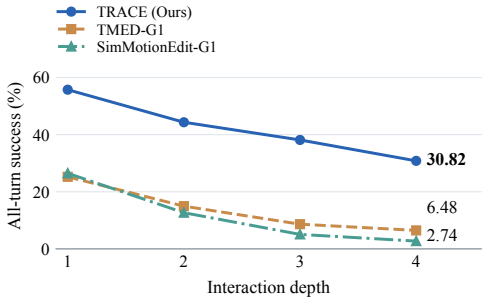


Figure 7: **Editing under shared masks.** All-turn success over 597 scripted chains and three sampling runs; higher is better.

Effect of scope-construction components. The full pipeline matches the prescribed masks on all 199 requests. With parser outputs fixed, whole-body and full-duration masks unnecessarily release unrelated coordinates or time intervals, while wrist-only masks exclude proximal arm joints needed for the requested change. These errors highlight the roles of body selection, temporal localization, and kinematic-chain expansion in constructing a compact editing scope. Figure 6 evaluates agreement with prescribed scopes; downstream editing quality is assessed separately.

5.3 Multi-Turn Editing Analysis

Multi-turn editing under shared preservation constraints. TRACE achieves the highest all-turn success at every evaluated depth. At four turns, success reaches 30.82%, compared with 6.48% for TMED-G1 and 2.74% for SimMotionEdit-G1. All 597 chains remain in the denominator, and each request is evaluated relative to the method's preceding output. Shared reference projection gives all methods zero protected-coordinate error, so the performance gap reflects their ability to realize successive changes under the same preservation constraint. Table 9 provides additional motion-quality diagnostics.

6 Conclusion and Limitations

TRACE combines motion-grounded typed-event conditioning, gravity-referenced diffusion, and source-relative editing for Unitree G1 motion. The complete generator achieves strong text–motion alignment and human ratings. Matched generation controls support the utility of event conditioning, realization-aligned auxiliary targets, and the evaluated kinematic design. Under shared scopes and reference projection, the complete editor achieves higher sustained completion of scripted hand-height requests than the adapted learned editors. The current evaluation is limited to text-conditioned G1 motion under flat-ground kinematics. Scene-grounded generation, broader editing requests, and longer-horizon interaction remain directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, implement methods, and interpret results. We used generative AI tools to generate synthetic datasets, assist with translation, clean or reformat datasets, or support qualitative and thematic data analysis. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to summarize or analyze existing literature, discover research topics or identify gaps, brainstorm, and revise the wording and organization of the manuscript. We have reviewed all AI-assisted work. For manuscript preparation, we reviewed AI-assisted drafts through iterative revision, checking that the wording and organization reflected the intended method. We also corrected the presentation of pseudocode, citation formatting, and figure captions to maintain consistency with the manuscript and figures. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J. Black, and Gül Varol. MotionFix: Text-driven 3D human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. URL https://motionfix.is.tue.mpg.de/.

Bones Studio. BONES-SEED: Skeletal everyday embodiment dataset. Dataset, n.d. URL https://huggingface.co/datasets/bones-studio/seed. Accessed September 21, 2026.

Jianuo Cao, Yuxin Chen, Yuzhen Song, Masayoshi Tomizuka, Chenran Li, and Thomas Tian. TEXEDO: Test time scaling for controller-aware language-conditioned humanoid motion generation. *arXiv preprint arXiv:2606.22998*, 2026. URL https://arxiv.org/abs/2606.22998.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024.

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

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbabu, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica K. Hodgins, Linxi Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.066.

Siqiao Huang, Kun-Ying Lee, Dongming Qiao, Guanqi He, Zhenyu Wang, Yitang Li, Shaoting Zhu, and Hang Zhao. OMG: Omni-modal motion generation for generalist humanoid control. *arXiv preprint arXiv:2606.10340*, 2026. URL https://arxiv.org/abs/2606.10340.

Haozhe Jia, Jianfei Song, Yuan Zhang, Honglei Jin, Youcheng Fan, Wenshuo Chen, Wei Zhang, and Yutao Yue. ECHO: Edge-cloud humanoid orchestration for language-to-motion control. *arXiv preprint arXiv:2603.16188*, 2026. URL https://arxiv.org/abs/2603.16188.

Yingzhao Jian, Zhongan Wang, Yi Yang, and Hehe Fan. Endowing gpt-4 with a humanoid body: Building the bridge between off-the-shelf vlms and the physical world. In *International Conference on Learning Representations*, volume 2026, pp. 134007–134065, 2026.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

Zhenyu Jiang, Yuqi Xie, Jinhan Li, Ye Yuan, Yifeng Zhu, and Yuke Zhu. Harmon: Whole-body motion generation of humanoid robots from language descriptions. *arXiv preprint arXiv:2410.12773*, 2024.

Zhe Li, Yangyang Wei, Boan Zhu, Yibo Peng, Tao Huang, Pengwei Wang, Zhongyuan Wang, Cheng Chi, Chang Xu, and Shanghang Zhang. From language to locomotion: Retargeting-free humanoid control via motion latent guidance. In *International Conference on Learning Representations*, volume 2026, pp. 129292–129319, 2026.

Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. Simmotionedit: Text-based human motion editing with motion similarity prediction. *arXiv preprint arXiv:2503.18211*, 2025. URL https://arxiv.org/abs/2503.18211.

Shaohui Lin, Zhenwu Shi, Jingyu Gong, Jiao Xie, Yu Zhou, Baochang Zhang, Lizhuang Ma, and Chia-Wen Lin. Spatial temporal synergy: Balancing change and invariance in text driven 3D human motion editing. *arXiv preprint arXiv:2608.16008*, 2026. URL https://arxiv.org/abs/2608.16008.

Runqi Ouyang, Haoyun Li, Zhenyuan Zhang, Xiaofeng Wang, Zeyu Zhang, Zheng Zhu, Guan Huang, Sirui Han, and Xingang Wang. Motion-R1: Enhancing motion generation with decomposed chain-of-thought and RL binding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7a4cf58610e1c911e06970c8d75b86dc-Abstract-Conference.html.

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

Yiyang Shao, Xiaoyu Huang, Bike Zhang, Qiayuan Liao, Yuman Gao, Yufeng Chi, Zhongyu Li, Sophia Shao, and Koushil Sreenath. Langwbc: Language-directed humanoid whole-body control via end-to-end learning. *arXiv preprint arXiv:2504.21738*, 2025.

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

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Bermano, and Michiel van de Panne. CLoSD: Closing the loop between simulation and diffusion for multi-task character control. In *International Conference on Learning Representations*, 2025.

Weiji Xie, Jiakun Zheng, Jinrui Han, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. TextOp: Real-time interactive text-driven humanoid robot motion generation and control. *arXiv preprint arXiv:2602.07439*, 2026. URL https://arxiv.org/abs/2602.07439.

Payam Jome Yazdian, Rachel Lagasse, Hamid Mohammadi, Eric Liu, Li Cheng, and Angelica Lim. MotionScript: Natural language descriptions for expressive 3D human motions. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 21574–21581. IEEE, 2025. doi: 10.1109/IROS60139.2025.11247385.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://nvlabs.github.io/PhysDiff/.

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

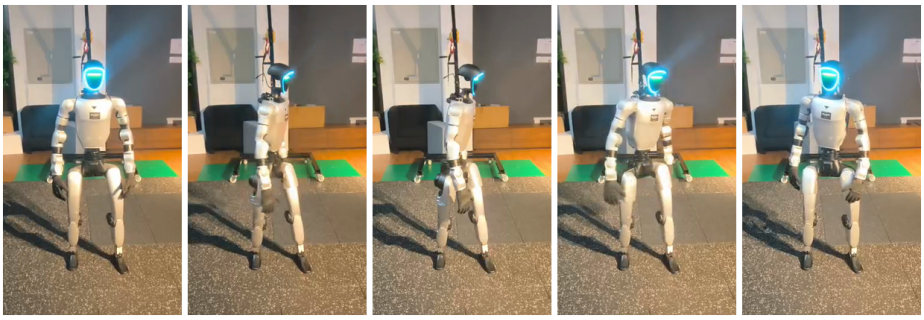
Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. FineMoGen: Fine-grained spatio-temporal motion generation and editing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 13981–13992, 2023b.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1551c01d7a3d0bf21e2518331e9f7074-Abstract-Conference.html.

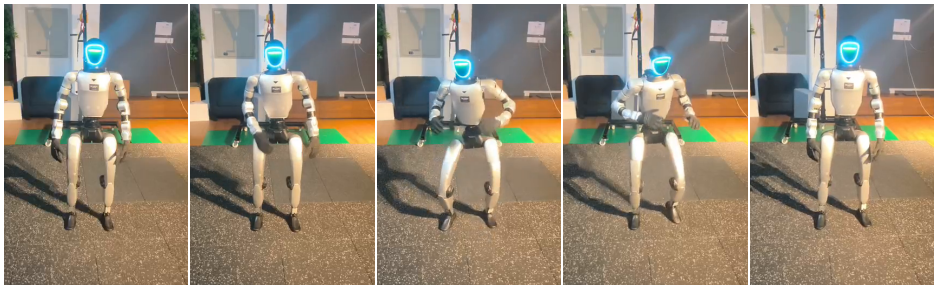
A REAL-ROBOT QUALITATIVE DEMONSTRATIONS

Figure 8 presents real-robot demonstrations of left–right looking, a hip-hop routine, and a weight-shifting groove.

(a) left–right looking "Can you slowly turn your head to the left and then to the right?"



(b) Hip-hop routine. "person quickly performs a hip-hop dance routine that includes a leg kick, arm strike, and return"



(c) Weight-shifting groove. "grooving by shifting bodyweight from one leg to another, hands on hips to hip hop music"

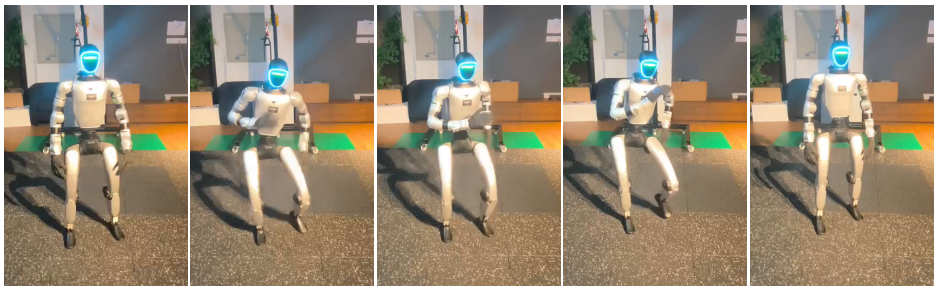


Figure 8: **Real-robot demonstrations.** Panel (a) corresponds to the in-the-wild request in Figure 3. Panels (b,c) sample 0/25/50/75/100% of each recording. Fixed spatial crops preserve poses and temporal order; clips are not time-synchronized.

B Sparse Plan Design and Validation

B.1 Motion-Derived Sparse Plan Construction

Clean-motion support, root, and palm signals yield confidence-weighted 16-channel targets on the four-frame patch grid. The Plan is sparse in vocabulary and dense in time; motion length is supplied separately.

Signals and coordinates. We use the clean G1 trajectory at 20 FPS, reconstructed from the Gravity-Gauged state. Leg FK evaluates four sole points per foot; palm FK follows the robot's hinge chains without clipping joint angles. The ground $\bar{z}$ is the fifth-percentile order statistic of the pooled minimum sole heights. Root height is measured above $\bar{z}$; hand height is the vertical palm displacement relative to the pelvis in the gravity-aligned frame, not a shoulder-joint distance.

Table 2: Sparse-event schema. Channels are zero-indexed. The four categorical groups use softmax; the feet are independent sigmoid probabilities. All distances are in meters.

Channels	Quantity	Fixed states / rules
0–1	Left/right support	Enter at height ≤ 0.035 and speed ≤ 0.20 m/s; leave at height > 0.055 or speed > 0.30 m/s.
2–4	Root vertical speed	Descending / level / ascending, split at $-0.08, +0.08$ m/s.
5–7	Root height	Low / middle / high, split at $0.60, 0.85$.
8–11	Left hand height	H0 / H1 / H2 / H3, split at $0, 0.25, 0.45$ relative to pelvis.
12–15	Right hand height	Same four bins as the left hand.

Temporal annotation and uncertainty. Support uses hysteresis. Frames between the enter/leave thresholds are uncertain, and contact runs shorter than three frames are confidence-masked. Lack of stable support is not an airborne label. For root/hand categories, we first apply a five-frame moving average with edge padding; root vertical speed is the gradient of smoothed height times 20. Frames within 0.025 m of a height threshold or 0.02 m/s of a speed threshold are uncertain. H3 is an operational height bin, despite its implementation label *above_shoulder*; it is not measured relative to an anatomical shoulder.

Each patch pools up to four valid frames. Its target is the histogram over confident frames and its confidence is the fraction of confident frames. No confident frame gives zero weight, not a negative event. For each categorical group g and class j,

$$Y_{\ell,g,j} = \frac{\sum_{t\in B_\ell} w_{t,g}\mathbf{1}[y_{t,g}=j]}{\max(1, \sum_{t\in B_\ell} w_{t,g})}, \qquad W_{\ell,g} = \frac{\sum_{t\in B_\ell} w_{t,g}}{|B_\ell|}. \tag{9}$$

The same confident-frame pooling applies to each binary foot label. The padded grid has 128 patches, of which $L = \lceil N/4 \rceil$ are valid. This representation does not use textual phases or learned phase boundaries.

Training and inference use. Training labels supervise the joint LLM planner, which is frozen during diffusion training. Equation 3 specifies the confidence fallback and 50/50 target/prediction mixture. Test conditions use only captions and predicted durations; the main model's duration-planner prose is not encoded as a condition. Training settings are detailed in Appendix E.2.

For example, patch 87 covers frames $[348, 352)$. Its root-height target is $(0.00, 0.67, 0.33)$ (low, middle, high), not just the displayed mode. Its group confidence is 0.75; the remaining weight falls back to the predicted distribution during target-conditioned training.

The example reproduces cached training targets: zero confidence means unknown, and the soft probabilities do not guarantee contact. These heuristic labels have no human semantic validation.

B.2 Main-Model Matched Plan Controls

The Aligned Plan checkpoint is exactly the model in Table 1. All three controls retain identical residual capacity, raw-caption features, 32k-step EMA initialization, optimizer, losses, and 8k continuation budget. They share the frozen joint event-and-duration LLM described in Appendix E.2.

Table 3: Actual cached training Plan example. Caption: *a simple jump in place*. $N = 512$ frames (25.6 s), $L = 128$ valid patches. The first non-mirrored jump UID in cache order is used. Patches show endpoints and first extrema of support, ascent, and confidence, selected solely from annotations, independently of model performance. Entries show support probability or categorical mode followed by confidence W.

Patch	L foot	R foot	Root speed	Root height	L hand	R hand
0	1.00/1.00	1.00/1.00	level/1.00	middle/1.00	H0/1.00	H0/1.00
4	0.00/0.50	0.00/0.50	up/1.00	high/0.75	H0/1.00	H0/1.00
87	0.67/0.75	0.50/1.00	up/1.00	middle/0.75	H0/1.00	H0/0.50
127	1.00/1.00	1.00/1.00	level/1.00	middle/1.00	H0/1.00	H0/1.00

Text Only zeros event inputs but retains trainable branch capacity. Global Events averages every channel over valid patches and broadcasts that mean at both training and inference. Aligned Plan preserves the patchwise sequence. Both event-conditioned variants use Equation 3. Thus event content and timing are the interventions, with no change to the underlying training budget.

All arms use the same 13,302 test UIDs, EMA, DDIM-50, CFG 2.5, and batch size 64. Table 4 reports absolute semantic results. These results do not measure independent-training variance. Targets enter training supervision only, never test conditions. The controls were developed after model selection and were not pre-registered.

Table 4: Matched LLM Plan controls on test. Aligned Plan is the main-table model. R@k is in percent.

Model	R@1↑	R@2↑	R@3↑	MM-Dist↓	FID↓	Diversity
Text Only	73.10	85.72	89.81	0.8341	0.0464	1.3817
Global Events	73.65	86.17	90.26	0.8310	0.0448	1.3809
Aligned Plan (Ours)	**73.84**	**86.24**	**90.31**	**0.8165**	**0.0434**	1.3766

Aligned Plan improves the displayed semantic criteria over both controls, consistent with patchwise events resolving temporal ambiguity left by global averages. Event prevalence can aid semantic conditioning without preserving when support and height changes should occur. These results support time-aligned conditioning but do not by themselves identify the preferred supervision source.

We additionally isolate the source of Plan supervision by replacing the motion-derived target in the 50% target-conditioned training branch with caption-derived events from a caption-only language model. The inference LLM and its predicted test Plan remain fixed, as do initialization, raw-caption features, residual capacity, 8k updates, objectives, and sampler. This is a supervision-source control, not a comparison of independently trained planners.

Table 5: Matched Plan-supervision source control on test. Both rows use the same predicted Plan at inference; only the training target source changes. R@k is in percent.

Supervision	R@1↑	R@2↑	R@3↑	MM-Dist↓	FID↓	Diversity
Caption-derived	73.26	85.89	90.00	0.8331	0.0458	1.3817
Motion-derived (Ours)	**73.84**	**86.24**	**90.31**	**0.8165**	**0.0434**	1.3766

Motion-derived supervision improves all reported retrieval and distributional criteria. A caption admits several valid support, root, and hand trajectories; selecting one from text alone can therefore disagree with the paired motion used as the diffusion target. Motion-derived auxiliary targets describe that same target trajectory, whereas caption-derived events may describe another plausible realization. The matched comparison supports the utility of this realization-aligned construction under the fixed inference planner; both variants still require text-only Plan prediction at test time.

C Kinematic Design and Objectives

We specify the state encoding and decoding, expand the grouped objective, and give the physical ablation settings and absolute values underlying Figure 5.

C.1 GRAVITY-GAUGED STATE: ENCODING AND DECODING

Coordinate convention and heading removal. Let $p_t \in \mathbb{R}^3$ and $R_t \in \mathrm{SO}(3)$ denote root position and the body-to-world rotation at motion frame $t = 0, \ldots, T-1$. The world z axis is vertical; the body's first axis is forward. Positions are stored in centimeters and joint angles in degrees, while the yaw angles below are in radians. Input root Euler angles are converted from degrees to radians; $(\phi_t, \vartheta_t, \psi_t^{\mathrm{Euler}})$ form $R_t = R_z(\psi_t^{\mathrm{Euler}})R_y(\vartheta_t)R_x(\phi_t)$. The stored translation triplet $(p^{(0)}, p^{(1)}, p^{(2)})$ is reordered as world $(x, y, z) = (p^{(2)}, p^{(0)}, p^{(1)})$ before encoding. Write e_x, e_y, e_z for the Cartesian unit vectors. Projecting the root's forward axis onto the horizontal plane gives

$$\begin{aligned} v_t &= (I - e_z e_z^\top) R_t e_x, & f_t &= \begin{cases} v_t/\|v_t\|_2, & \|v_t\|_2 > 10^{-8}, \\ e_x, & \text{otherwise}, \end{cases} \\ H_t &= [f_t,\ e_z \times f_t,\ e_z] = R_z(\psi_t), \quad \psi_t = \mathrm{atan2}(f_t^y, f_t^x). \end{aligned} \tag{10}$$

The yaw-free residual is $Q_t = H_t^\top R_t$. Its continuous six-coordinate encoding stores the *first two columns*, concatenated columnwise:

$$\mathbf{r}_t^{\mathrm{tilt}} = \begin{bmatrix} Q_t e_x \\ Q_t e_y \end{bmatrix} \in \mathbb{R}^6. \tag{11}$$

Thus tilt is an absolute per-frame residual rotation, not a pair of Euler angles, a gravity-vector measurement, or a frame-to-frame rotation increment.

Translation and yaw increments. For $t < T-1$, the remaining root coordinates are

$$\begin{aligned} (\Delta p_t^x, \Delta p_t^y) &= \left[H_t^\top (p_{t+1} - p_t)\right]_{xy}, & h_t &= p_t^z, \\ \Delta\psi_t &= \mathrm{atan2}(\sin(\psi_{t+1} - \psi_t), \cos(\psi_{t+1} - \psi_t)), & (s_t, c_t) &= (\sin\Delta\psi_t, \cos\Delta\psi_t). \end{aligned} \tag{12}$$

These are displacements per frame, not velocities; no frame-rate factor enters their integration. Together with h_t, the six tilt coordinates and 29 joint angles, they give the 40D state in Equation 4. The encoder copies the penultimate planar and yaw increments into the last frame; those last increments are never integrated. For a single frame, planar increments are zero and $(s_0, c_0) = (0, 1)$.

Decoding and the initial gauge. Split a decoded tilt vector into $(a_{1,t}, a_{2,t}) \in \mathbb{R}^3 \times \mathbb{R}^3$. The implementation uses Gram–Schmidt normalization:

$$\begin{aligned} \mathcal{N}_\epsilon(a) &= a/\max(\|a\|_2, \epsilon), \\ b_{1,t} &= \mathcal{N}_\epsilon(a_{1,t}), \qquad b_{2,t} = \mathcal{N}_\epsilon\big(a_{2,t} - (b_{1,t}^\top a_{2,t}) b_{1,t}\big), \\ \widetilde{Q}_t &= [b_{1,t},\ b_{2,t},\ b_{1,t} \times b_{2,t}]. \end{aligned} \tag{13}$$

Here $\epsilon = 10^{-6}$ in the differentiable FP32 Torch decoder and 10^{-12} in the FP64 NumPy export decoder. The result is orthonormal when both normalized vectors have pre-normalization norms at least ϵ; the clamp alone does not repair zero or collinear pairs.

For predicted (s_t, c_t), recover $\widetilde{\Delta\psi}_t = \mathrm{atan2}(s_t, c_t)$. The Torch decoder substitutes $(0, 1)$ if the pair norm is at most 10^{-6}; the NumPy export uses the pair directly. Starting with $\widetilde{\psi}_0 = 0$ and $\widetilde{p}_0^{xy} = 0$, decode

$$\begin{aligned} \widetilde{\psi}_t &= \sum_{i=0}^{t-1} \widetilde{\Delta\psi}_i, & \widetilde{H}_t &= R_z(\widetilde{\psi}_t), \\ \widetilde{p}_t^{xy} &= \sum_{i=0}^{t-1} \left[\widetilde{H}_i(\Delta p_i^x, \Delta p_i^y, 0)^\top\right]_{xy}, \quad \widetilde{p}_t^z = h_t, \\ \widetilde{R}_t &= \widetilde{H}_t \widetilde{Q}_t. \end{aligned} \tag{14}$$

The 29 joint angles are copied unchanged. Initial global planar position and yaw are intentionally omitted: for an exactly encoded trajectory, $\widetilde{R}_t = R_z(-\psi_0)R_t$. Restoring the original global frame requires its initial planar translation and yaw as external metadata. Height is retained, not integrated or reset to zero.

Numerical input and masking. The denoiser operates on featurewise normalized states $z_t = (x_t - \mu)/\sigma$, with training-only statistics and a standard deviation floor of 10^{-4}. Decoding first applies $x_t = \sigma \odot z_t + \mu$. In a padded batch, each sequence has a valid prefix; increments from padded frames are zeroed and padded outputs do not enter any objective. The same canonical decoder is applied separately to the prediction and clean target for the geometric losses below.

C.2 Generation Objective Groups and Contact Implementation

Objective groups. The compact objective in Section 3.3 collects the component losses into three groups, with component-specific weights:

$$\mathcal{L}_{\text{motion}} = \mathcal{L}_{\text{rec}} + \lambda_v \mathcal{L}_{\text{vel}} + \lambda_a \mathcal{L}_{\text{acc}}, \tag{15}$$

$$\mathcal{L}_{\text{geometry}} = \lambda_p \mathcal{L}_{\text{int-pos}} + \lambda_\psi \mathcal{L}_{\text{int-yaw}} + \lambda_g \mathcal{L}_{\text{gravity}} + \lambda_u \mathcal{L}_{\text{yaw-unit}}, \tag{16}$$

$$\mathcal{L}_{\text{contact}} = \lambda_s \mathcal{L}_{\text{slip}} + \lambda_b \mathcal{L}_{\text{anchor}} + \lambda_h \mathcal{L}_{\text{height}} + \lambda_f \mathcal{L}_{\text{penetration}}. \tag{17}$$

Motion losses supervise normalized clean states and temporal derivatives; geometry losses supervise integrated position/yaw, gravity, and yaw-unit norm. Contact terms use decoded FK. Component activation is specified in Appendix C.3.

Geometric inputs and reduction. At a sampled diffusion noise level, the denoiser predicts a normalized *clean* sequence $\widehat{z}_{b,t}$, with clean target $z_{b,t}$. Both are denormalized using the same training statistics and decoded by Appendix C.1, yielding $(\widehat{p}_{b,t}, \widehat{\psi}_{b,t}, \widehat{R}_{b,t})$ and $(p_{b,t}, \psi_{b,t}, R_{b,t})$. Each trajectory integrates its own predicted or target increments from zero initial planar position and yaw; there is no additional rigid alignment or use of target headings to integrate predicted translations. These losses supervise the one-step clean prediction, not a completed reverse-diffusion rollout. Training does not clip this prediction before computing geometry.

Let $m_{b,t} \in \{0, 1\}$ be the valid-frame mask and $V = \sum_{b,t} m_{b,t} > 0$. Define body-up vectors $\widehat{u}_{b,t} = \widehat{R}_{b,t} e_z$ and $u_{b,t} = R_{b,t} e_z$. Let $(\widehat{s}_{b,t}, \widehat{c}_{b,t})$ be the two denormalized yaw coordinates *before* the decoder's small-norm fallback. The implemented geometry terms are

$$\mathcal{L}_{\text{int-pos}} = \frac{1}{3V} \sum_{b,t} m_{b,t} \left\| \frac{\widehat{p}_{b,t} - p_{b,t}}{s_p} \right\|_2^2, \qquad s_p = 100, \tag{18}$$

$$\mathcal{L}_{\text{int-yaw}} = \frac{1}{V} \sum_{b,t} m_{b,t} \left[1 - \cos(\widehat{\psi}_{b,t} - \psi_{b,t}) \right], \tag{19}$$

$$\mathcal{L}_{\text{gravity}} = \frac{1}{3V} \sum_{b,t} m_{b,t} \|\widehat{u}_{b,t} - u_{b,t}\|_2^2, \tag{20}$$

$$\mathcal{L}_{\text{yaw-unit}} = \frac{1}{V} \sum_{b,t} m_{b,t} \left(\sqrt{\widehat{s}_{b,t}^2 + \widehat{c}_{b,t}^2} - 1 \right)^2. \tag{21}$$

In Equation 18, p is the full three-coordinate root position: its planar coordinates are integrated, while its height is predicted directly. The factor $s_p = 100$ converts centimeter-valued errors to meters before squaring; division by $3V$ averages coordinates as well as valid frames. This is not a planar-only, terminal-position, or per-sequence-normalized loss. Cumulative yaw is measured in radians, and its cosine penalty respects angle periodicity without a degree conversion or division by π.

The gravity term matches the *target body-up direction*, not a fixed upright pose. It uses the third column of the complete decoded rotation, including heading composition, and therefore retains demonstrated leaning. The yaw-unit term has scalar target one; it penalizes $(\|v\|_2 - 1)^2$, not $(\|v\|_2^2 - 1)^2$. The latter three terms are dimensionless. All sums pool valid frames across the batch, so longer sequences contribute more frames; there is no additional timestep-dependent reweighting. Empty masks contribute zero, with implementation denominators clamped below by one. Denormalization and geometric computations use FP32. The reported geometry weights are $(\lambda_p, \lambda_\psi, \lambda_g, \lambda_u) = (0.25, 0.10, 0.25, 0.02)$.

Contact penalties. Let $p^f_{t,k}$ be sole point k of foot f, $z^f_t = \min_k p^{f,z}_{t,k}$, and c^f_t the clean-motion support mask. Define uninterrupted-contact weights $w^f_{t,\delta} = \prod_{j=0}^{\delta} c^f_{t+j}$. With masked means over valid batch, time, feet, and applicable point/coordinate entries, the implemented penalties are

$$\mathcal{L}_{\text{slip}} = \left\langle \rho_{\beta_v}\left(20(p^{f,xy}_{t+1,k} - p^{f,xy}_{t,k})\right)\right\rangle_{w^f_{t,1}}, \tag{22}$$

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\mathcal{D}|}\sum_{\delta\in\mathcal{D}} \left\langle \rho_{\beta_p}(p^{f,xy}_{t+\delta,k} - p^{f,xy}_{t,k})\right\rangle_{w^f_{t,\delta}}, \tag{23}$$

$$\mathcal{L}_{\text{height}} = \left\langle \rho_{\beta_h}(z^f_t - \bar{z}_{\text{gt}})\right\rangle_{c^f_t}, \tag{24}$$

where $\mathcal{D} = \{2, 4, 8\}$ excludes lags not shorter than the sequence. Position and velocity use meters and meters per second, respectively.

The reported main checkpoint uses differentiable FK for both six-joint G1 legs. Each ankle-roll link carries four sole sample points, in meters: $(-0.05, \pm 0.025, -0.035)$ and $(0.12, \pm 0.03, -0.035)$. The ground estimate pools the clean minimum sole height of each foot over valid frames and selects the fifth-percentile order statistic. A candidate support frame has sole height at most 0.035 m above this ground and foot-center planar speed at most 0.20 m/s. Singleton contacts are removed by requiring an adjacent contact frame. All masks are computed from clean motion without gradient propagation; generated motion is used only in the loss arguments.

Slip uses adjacent support frames; anchors require uninterrupted support at lags 2/4/8 (0.1/0.2/0.4 s), normalized separately. Height uses the lowest sole point, not the ankle origin. Invalid frames are excluded; empty masks contribute zero. The elementwise robust penalty is

$$\rho_\beta(r) = \begin{cases} r^2/(2\beta), & |r| < \beta, \\ |r| - \beta/2, & \text{otherwise.} \end{cases} \tag{25}$$

Its scales are $\beta_v = 0.05$ m/s for slip and $\beta_p = \beta_h = 0.01$ m for anchor and height. All three loss weights are 0.05 in the reported checkpoint. These soft penalties do not guarantee zero sliding; support-only height does not penalize swing-frame penetration.

Penetration and Plan-mask distinction. The current continuation and the final physical-control bundle also use

$$\mathcal{L}_{\text{penetration}} = \left\langle \max(0, \bar{z}_{\text{gt}} - \widehat{z}^f_t)\right\rangle_{\text{valid}}, \tag{26}$$

with weight one, zero tolerance, and FP32 contact computation. This hinge averages over valid frames and feet, independently of support.

Loss masks use the clean-motion selector above, independently of predicted Plan probabilities. Plan annotation instead uses hysteresis, confidence, three-frame runs, and patch pooling (Appendix B.1). Generated motion cannot avoid a loss by declaring itself out of contact.

C.3 Motion-Consistency Ablation Protocol

All four controls use raw captions and the current sparse-event condition, including the same 50/50 training mixture and frozen joint duration/event LLM. They share training data, global batch 256, 32k updates from scratch, learning rate 10^{-4}, 500-step warm-up, and EMA 0.999. Testing uses the same predicted events, DDIM-50, CFG 2.5, and 13,302 complete test motions.

All variants retain reconstruction, velocity, and acceleration losses. Gravity-Gauged core has no geometry/contact objectives. Geometry adds integrated-position, integrated-yaw, gravity, and yaw-unit losses with weights $(0.25, 0.10, 0.25, 0.02)$. Contact adds slip/anchor/height with weight 0.05 each and penetration with weight one; the final step tests this bundle, not individual foot losses. Heading-local and Gravity-Gauged retain their native 38D/40D targets, normalization, and decoders; this is a matched-budget representation comparison, not a one-coordinate intervention.

For variant v and metric m, the plotted ordinate is

$$\widetilde{E}_{v,m} = \frac{\text{Mean}(E_{v,m})}{\text{Mean}(E_{\text{Heading},m})}. \tag{27}$$

Table 6: Absolute physical errors underlying Figure 5. Mean $\pm$ SD over repeated sampling, in cm; lower is better. Sliding and Drift measure planar endpoint displacement over reference contact runs, using ankle and sole-point geometry, respectively.

Variant	Sliding	Drift	Height error	Penetration
Heading-local	28.152 ± 0.090	31.756 ± 0.074	3.385 ± 0.018	0.2659 ± 0.0004
Gravity-Gauged	26.047 ± 0.010	29.558 ± 0.023	2.596 ± 0.003	0.1575 ± 0.0021
+ Geometry	24.456 ± 0.033	27.817 ± 0.055	2.533 ± 0.008	0.1970 ± 0.0017
+ Contact	23.047 ± 0.062	26.383 ± 0.068	2.366 ± 0.007	0.1333 ± 0.0017

Each metric has its own unrounded denominator; no aggregate score is formed. Geometry increases penetration before Contact reduces it, so improvement is not monotonic for every metric.

D Automatic Masking and Motion Editing

The request interface and editor training precede the current mask/multi-turn protocol and same-cohort motion-quality comparison.

D.1 Language Interfaces and Qualitative Whole-Body Example

This section specifies the language interfaces and contextualizes the full-body refinement in Figure 1. The two language adapters share the instruction-tuned backbone reported in Section 4.1, but are trained and serialized independently because generation and editing expose different output schemas. Both adapters are frozen when their respective diffusion models are trained or sampled.

Adapter interfaces. The generation adapter receives a standalone instruction and emits a duration plus a sequence of typed event intervals. Each interval records its start and end, left/right support, root vertical state, root-height state, and left/right hand-height states. Deterministic validation clips interval endpoints to the predicted duration, resolves overlaps by output order, fills uncovered regions with unknown states, and rasterizes the result onto the four-frame Plan grid. The original instruction is still encoded separately; the serialized Plan is an additional condition, not a replacement for raw text.

The editing adapter receives the follow-up instruction together with the duration and a source-only kinematic summary. It emits an attribute, effector, signed change, and temporal interval. A deterministic compiler maps this record to a body–time scope, a source-relative target when the attribute has a metric interpretation, and revised source-event channels when the request changes an event-level property. Invalid or incomplete records produce a closed scope and leave the source unchanged.

Table 7: Serialized language interfaces. Both outputs are schema-validated before being consumed by a motion model.

Adapter	Supervised output	Saved inference artifact
Generation	duration and typed support/root/hand intervals	adapter weights, tokenizer/template identifier, schema, and event vocabulary
Editing	attribute, effector, signed change, and interval	adapter weights, tokenizer/template identifier, schema, and compiler vocabulary

Training examples are stored as an instruction followed by one canonical JSON object. Generation targets come from the motion-derived annotation in Appendix B.1; editing targets pair paraphrased requests with typed deltas and compiler scopes. The reproducible checkpoint for each adapter consists of its parameter-efficient weight update, the base-model revision, prompt template, tokenizer files, output schema, and normalization or vocabulary metadata. Decoding is constrained to one JSON object and followed by the deterministic validator above. No conversational history is required beyond the source state and the self-contained follow-up request.

Whole-body refinement in Figure 1. Figure 1 presents a jumping-jack refinement with a deeper preparatory crouch and a higher jump. It involves coordinated root and lower-body changes and illustrates a broader editing scenario than the controlled hand-height task. The two-wrist target

formulation and quantitative success criteria below apply to the hand-height protocol. The figure is reported separately as a qualitative whole-body case and is not included in the quantitative editing cohort.

Quantitative boundary. The reported editing benchmark starts from held-out G1 source motions so that requested changes and protected coordinates can be measured against a common source. After the first edit, each subsequent turn consumes the editor's own preceding output, and therefore evaluates multi-turn accumulation. It does not measure the additional failure mode introduced when the initial source is first generated by TRACE. The Figure 1 pipeline demonstrates that end-to-end interface qualitatively.

D.2 Markov Editing State and Relative Requests

For the supported self-contained requests, the state and transition follow Equation 7: $X_t = (M_t, u_t)$ and $X_{t+1} = (M_{t+1}, u_{t+1})$, with external next utterance u_{t+1}. The frozen parser emits a relative action; supervised editing and deterministic projection realize it without RL. The editor retains source duration and derives events from the source, not the generator's text predictions or an edited target.

Source-only automatic mask construction. Inputs are the request, frame count, and four-frame mean left/right wrist heights relative to the pelvis, computed by source FK. The supervised adapter predicts JSON fields for attribute, effector, signed offset, and interval; the compiler validates them and resolves supported time syntax. For hand $e^\star$, offset d, and interval $[a, b)$, compilation gives

$$S_{f,j} = \mathbf{1}[j \in \mathcal{J}(e^\star)]\mathbf{1}[a \le f < b], \tag{28}$$

$$D_{f,h} = \mathbf{1}[h = e^\star]\, d \sin^2\left(\frac{\pi(f-a)}{b-a-1}\right), \quad a \le f < b, \tag{29}$$

where $\mathcal{J}$ contains the seven arm coordinates; $D = 0$ outside the interval. The complement, including the selected arm outside its window, is protected. This fixed envelope defines the synthetic task; the parser does not predict a dense mask or access target/evaluation masks.

Offsets satisfy $0 < |d| \le 0.10$ m. Negated, ambiguous, or unsupported expressions leave the source unchanged.

D.3 Reference-Conditioned Editor: Architecture and Training

Training data. Retargeted official MotionFix pairs provide 5,387 training and 330 validation pairs. Local supervision adds 3,690 pairs from 256 training sources, six windows, both hands, signed 3/5/7 cm offsets, and six instruction templates. Source-only bounded IK constructs training targets; rejected candidates are recorded before evaluation. Selection uses 64 original-validation and 40 expanded-development cases. This training construction is separate from test-time inference, which uses no IK. Bounded IK is used offline to construct local training targets. At inference, the learned editor produces revisions through conditional denoising and reference projection, without per-request IK optimization or post-hoc smoothing.

Architecture. The independent editor uses 40D states, four-frame patches, at most 104 frames, an eight-layer width-448 denoiser with eight heads, and a two-layer source encoder. Text queries attend to source patches. Scope and two-wrist offset patches have separate zero-initialized projections; offsets are scaled by 5 cm. Its weights and normalization are independent of the main-table generator.

Original-edit training uses batch 192 and learning rate 10^{-4}. Relative-goal and event-aware continuations use batch 128, learning rates 5×10^{-5} and 2×10^{-5}, AdamW, EMA 0.99, a 300-step cosine schedule, and clean-state prediction. Selection maximizes mean joint success on the two validation sets, breaking ties by height RMSE at steps 2k/4k. Both TRACE continuations select step 4k.

Table 8: TRACE editor training stages. Mixed batches contain equal numbers of local and original MotionFix pairs.

Stage	Updates	Supervision / added condition
Original editing	8k	Original retargeted motion pairs
Mixed editing	8k	Local/original source+scope initialization
Relative goal	4k	Relative offsets and wrist goal loss
Event-aware	4k	Source-derived event residual and consistency

Local objectives. Let $\mathcal{L}_{\text{wrist}}$ denote wrist FK squared error normalized by $(0.05\,\text{m})^2$. The local reconstruction and goal groups are

$$\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{rec,edit}} + 0.2\mathcal{L}_{\text{vel,edit}} + 0.2\mathcal{L}_{\text{wrist}}, \tag{30}$$

$$\mathcal{L}_{\text{goal}} = 0.5\mathcal{L}_{\text{vertical}} + 0.1\mathcal{L}_{\text{horizontal}}. \tag{31}$$

The goal uses vertical and mean horizontal squared wrist errors, also normalized by $(0.05\,\text{m})^2$, against the source trajectory plus the public offset over editable entries. Original-pair losses retain reconstruction, velocity, and applicable auxiliary heads. Adapted baselines also receive the local goal term.

Source-derived events. The final stage adds a zero-initialized MLP over 16 source-motion event channels and four progress features. Only requested hand-height groups are revised by the relative offset; no target motion supplies inference events. The residual is attenuated during the last quarter of reverse denoising. Confidence-weighted hand-category cross entropy has weight 0.05. Current test cohorts and baseline checkpoints follow in Appendix D.4; same-cohort quality diagnostics are in Appendix D.5.

D.4 Automatic Localization and Sequential-Editing Protocols

Cohort and separation of questions. Of 200 frozen BONES-SEED/G1 candidates, 199 pass human review for naturalness, scope, and feasibility, independently of model outputs. Sources contain 64–104 frames; initial requests select one hand. Mirrored duplicates, training-source groups, and exact source-array hashes from earlier editing train/development sets are excluded; this does not establish semantic near-duplicate absence. Initial requests were reviewed; scripted chains were not independently reviewed. Mask-only scoring and shared-scope execution answer separate questions.

Mask-only structural interventions. We reuse the same frozen, fine-tuned instruction-parser predictions, whose public inputs contain the instruction, length, and source summary. Private reference masks are accessed only for scoring. Removing body selection opens all 40 coordinates in the predicted interval; removing temporal localization opens the selected arm throughout the valid sequence. Removing arm-chain expansion retains only the three wrist coordinates, compared with seven arm coordinates in the reference. No model is retrained and no threshold is tuned in this diagnostic.

The whole-body overlap follows from the ratio of reference arm coordinates to all state coordinates; wrist-only overlap follows from the retained fraction of the arm chain. Full-duration overlap reflects the average requested interval fraction. These geometric ratios explain the structural controls in Figure 6 without measuring motion quality. They do not establish a uniquely minimal physical scope.

Four-turn construction and inference. Each source generates three fixed chains. Let A be the original selected hand, B the other hand, and d its signed 5 cm offset. The same-arm pattern is $(A{+}d, A{+}d, A{-}d, A{-}d)$; alternating arms uses $(A{+}d, B{+}d, A{-}d, B{-}d)$; reversals use $(A{+}d, A{-}d, A{+}d, A{-}d)$. All methods receive the same explicit instructions, frame intervals, deterministic compiler, and per-step reference projection. Each turn uses the actual preceding generated motion, with no reset, teacher forcing, early termination, or failure exclusion. The source duration is unchanged. DDIM-50, CFG 1.5, EMA, batch 64, and paired sampling yield 21,492 outputs across methods, turns, and chains. Neither IK nor smoothing is applied to generated outputs.

Models and fairness boundary. Validation selects TRACE's event-aware step 4k and the mixed-training goal-loss checkpoints at steps 4k/2k for TMED-G1/SimMotionEdit-G1, before testing. TRACE retains relative offsets and source events; baselines retain text/source interfaces. Scope and sampling are shared, but architectures, prior training, and total FLOPs differ. These are local G1 adaptations, not published native benchmark scores.

Aggregation, feasibility, and uncertainty. The primary denominator includes all 597 chains from 199 sources; no chain is removed using solver-derived feasibility. At depth d, all-turn success requires joint edit-and-preserve success at every preceding turn, as defined in Appendix E.4. We average repeated samples within each chain, then bootstrap 199 source clusters while retaining all three patterns. Four-turn success has 95% source-cluster intervals of $[26.07, 35.62]\%$ for TRACE, $[4.13, 9.10]\%$ for TMED-G1, and $[1.23, 4.47]\%$ for SimMotionEdit-G1. These describe source-sampling uncertainty, not independent-training variability. Protected-coordinate error is zero for all methods under the shared projection, so it cannot explain a cross-method advantage.

D.5 Current-Cohort Editing Quality

We evaluate the same 199 sources and all 597 four-turn chains used in Figure 7. Results aggregate paired sampling runs. Each turn starts from the method's actual preceding output. No private target, feasibility filtering, or test-specific tuning is used.

Table 9: Current-cohort editing comparison. Success requires all four turns to pass. Quality diagnostics concern the fourth edit relative to its incoming motion. Lower is better for all quality columns.

Method	Success (%)	XY RMSE (cm)	Δv (deg/s)	Δa (deg/s^2)	New limits (%)
SimMotionEdit-G1	2.74	2.238	34.273	602.284	0.033
TMED-G1	6.48	2.206	52.730	967.980	0.043
TRACE	30.82	2.047	26.304	520.346	0.017

XY RMSE measures unintended horizontal wrist displacement over the active edit. Boundary diagnostics measure RMS velocity/acceleration changes relative to the incoming motion on the edited arm. New-limit rate counts editable coordinates that newly violate joint limits; pre-existing violations are excluded from this count. These diagnostics do not measure whole-body naturalness.

TRACE improves success and these quality diagnostics over the two learned adaptations. The controlled task does not establish whole-body editing quality; broader coordinated editing requires separate evidence.

E Data and Reproducibility Details

E.1 BONES-SEED Data Processing

Source and scope. We use BONES-SEED's supplied G1 retargeting and captions (Bones Studio, n.d.). Its 142,220 metadata entries include original and mirrored motions. We select a predominantly ground-supported subset and construct grouped splits; this is neither new motion capture nor the provider's official split.

Selection and filtering. Requiring body-position `standing` and temporarily excluding mirrors gives 61,292 originals. We reject malformed paths, Stunts, and metadata indicating non-standing or externally supported motion. At an intermediate 50 FPS, a kinematic screen rejects clips whose two lowest foot-collision points stay above world ground by 5 cm for at least one second; sampled rejection candidates are checked frame by frame. This leaves 58,574 originals.

A semantic filter removes 3,115 originals (1,969 keyword rules; 1,146 Qwen-Max classifications), leaving 55,459. It excludes difficult acrobatics and externally supported actions, retaining walking, running, turns, gestures, squats, daily dance, and small jumps. Persisted filtering decisions are reused. This classification is separate from deterministic Plan annotation and does not certify dynamic stability.

Mirrors and captions. Restoring 55,423 supplied mirrors gives 110,882 clips; 36 originals remain unpaired. No mirrors or captions are synthesized. We retain four natural, one technical, and two short-description fields, plus filename-matched temporal annotations. Mirror annotations come from their own records without left/right substitution; clip-level captions may still be shared. After whitespace/case-insensitive deduplication, the generator takes the first nonempty caption in this order; semantic evaluation uses the corresponding first natural description.

Grouped split and leakage checks. We connect records sharing a motion family, its mirror-canonical family, an original/mirror pair, or a normalized exact caption. Caption matching uses all available English alternatives and an existing locally translated Chinese metadata copy; the latter is not an additional motion source or an assertion of official Chinese annotation. Text normalization uses Unicode normalization, case folding, and punctuation/whitespace cleanup. Connected components are kept intact when assigning train/validation/test with seed 20260822 and target proportions 76/12/12, balancing clips and frames. Adding mirrors rebuilds the assignment and produces 2,557 final components. The recorded audit finds zero cross-split overlap for paths, exact and mirror-canonical families, caption keys, components, and mirror pairs. This is not an actor-disjoint split or a guarantee that every semantic paraphrase is disjoint; generic temporal annotation phrases are not grouping keys.

Table 10: Processed BONES-SEED counts. Full clips, packed windows, and eligible model examples are different units. Eligible windows have at least ten frames; durations count full clips at 20 FPS.

Split	Originals	Mirrors	Full clips	Packed / eligible windows	Hours
Train	42,156	42,131	84,287	86,451 / 86,385	170.54
Validation	6,650	6,643	13,293	13,803 / 13,799	26.95
Test	6,653	6,649	13,302	13,722 / 13,706	26.91

Temporal and geometric representation. The final 20 FPS trajectories are obtained directly from the 120 FPS source: output frame k copies source frame $6k$, without interpolation, smoothing, or IK. Frame counts and durations are recomputed at the target rate. Parsing retains root translation/orientation and the 29 G1 joint angles; unused non-root translation fields do not enter the model. We then convert to the 40D Gravity-Gauged state defined in Section 3.3: planar increments, absolute root height, yaw increments, a six-coordinate yaw-free tilt rotation, and joint angles. Appendix C.1 defines heading removal, columnwise tilt encoding, canonical root reconstruction, and terminal-frame handling. Means and standard deviations are estimated from training frames only, with a standard deviation floor of 10^{-4}; validation and test reuse these statistics.

Windowing and Plan supervision. Long clips are split into consecutive, nonoverlapping windows of at most 512 frames. Original/mirror pairs share window boundaries, and each window retains its parent motion ID, source offsets, and ordering. Temporal annotations are clipped and rebased to each window. Windows shorter than ten frames are excluded from the model's input manifests (66/4/16 in train/validation/test); padded frames are masked out of losses. Four-frame patches are formed only after these operations. The sparse-event Plan is derived from clean G1 motion through FK, thresholding, confidence masking, and patch aggregation, not copied from the provider's prose phase labels or freely written by an LLM. Appendix B.1 specifies every event channel and its thresholds. Training-motion labels supervise the joint LLM planner; test conditioning uses predicted events and durations, never target event labels.

Full-motion evaluation. The 13,706 eligible test windows are generated, decoded, and reassembled in parent-window order into exactly 13,302 full-motion IDs. The variable-length assembly concatenates heading-local root increments and joint states; it does not pad or truncate predictions to GT lengths and does not fill excluded short tails with GT frames. Each parent is scored once. Source-derived segmentation is retained, so the protocol predicts window durations but is not an unconstrained full-clip duration benchmark. This distinction explains why a segment count must not be reported as the number of independent test motions.

E.2 Training and Evaluation Configuration

Generation backbone and continuation. The frozen encoder in Section 4.1 supplies a 2,560-dimensional caption vector. The generator has 32,508,520 parameters, including residual branches but excluding upstream models, and feed-forward width 1,792. Training predicts clean states with a 1,000-step cosine schedule.

Table 11: Generation training schedule. Continuations start from the same raw-text EMA initialization and retain equal residual capacity. Physical from-scratch controls are specified in Appendix C.3.

Setting	Raw-text backbone	Sparse-event continuation
Updates	32k	8k
Global batch	256	128
Learning rate	10^{-4}	2×10^{-5}

Continuation uses 250-step warm-up and EMA 0.999. Contact losses use FP32, with weights in Appendix C.2. The event residual is scaled by $\min(1, k/[0.25(K-1)])$, $K = 1000$. Main-table inference uses EMA, DDIM-50, CFG 2.5, and batch 64.

Joint LLM planner. A supervised language-model adapter predicts duration and typed support, root, and hand events in a structured Plan. The event intervals are rasterized onto the four-frame patch grid; no separate support or root/hand prediction network is used in this configuration. The LLM is frozen for diffusion training and inference. The generator uses all 86,385 training segments with the joint 50/50 condition mixture in Equation 3; event extraction is defined once in Appendix B.1. Most predicted training conditions are in-sample, without out-of-fold separation. Test inputs contain no target events or durations. The Plan controls in Appendix B.2 reuse the same joint LLM frontend. Invalid Plans remain in the evaluation with target-free fallback: unknown support values are 0.5 and unknown categorical groups use uniform probabilities.

Retrained baselines. MDM, MoMask, and MotionGPT3 consume raw captions and predict heading-local 38D G1 motion. Their budgets are 64k steps (MDM), 17k VQ-autoencoder plus 50k base/residual Transformer steps (MoMask), and 50k motion-VAE plus 50k language-model steps with pretrained language initialization (MotionGPT3). OMG, ECHO, TextOP, and TeXedo retain official weights and generation settings.

Evaluation units. Main-test generation reassembles 13,706 eligible segments into 13,302 parent IDs, each scored once. Physical metrics and distinct uncertainty estimates are defined in Appendix E.4; repeated samples with fixed weights do not measure training variance.

E.3 Inference Procedure

Algorithm 1 distinguishes the main generation network F_θ from the independent editor E_η. Editing uses source features in addition to reference projection; generation predicts Sparse events from text. Neither branch receives an edited target motion.

E.4 Metric Definitions and Aggregation

E.4.1 Semantic and Distributional Metrics

All main-table methods use one frozen G1 evaluator and 13,302 parent IDs. The text branch uses GloVe/POS features and a bidirectional GRU; the motion branch uses normalized 38D states, temporal convolutions, and a bidirectional GRU. Both yield unit-normalized 512D embeddings. Training uses a symmetric contrastive objective on the training split, with no generated-test updates. Let e_i, z_i, and z_i^{gt} be paired text, generated-motion, and target-motion embeddings. Distances are Euclidean.

R-Precision (R@1/R@2/R@3, higher is better). For every method, retrieval is motion-to-text: z_i ranks its caption among 32 candidates, including 31 others sampled without replacement. For

Algorithm 1 Sparse generation and independent relative editing

Require: instruction u, optional source motion M

$e \leftarrow f(u)$
if $M = \varnothing$ **then**
  $(N, C) \leftarrow \textsc{Rasterize}(\textsc{LLMPlan}(u))$; $x_K \sim \mathcal{N}(0, I)$
else
  $N \leftarrow \text{length}(M)$; $\delta \leftarrow \textsc{ParseLLM}(u, N, s(M))$
  $(S, D) \leftarrow \textsc{ValidateAndCompile}(\delta, u, N)$
  if parsing or compilation fails **then**
    return M
  end if
  $m \leftarrow \text{Enc}(M)$
  $x_K \leftarrow S \odot \epsilon + (1 - S) \odot q_K(m)$
end if
for each reverse DDIM timestep k **do**
  if $M = \varnothing$ **then**
    $\widehat{x}_0 \leftarrow F_\theta(x_k, k, e, C, N)$ with CFG
  else
    $\widehat{x}_0 \leftarrow E_\eta(x_k, k, e, m, S, D)$ with CFG
  end if
  $\widetilde{x}_{k-1} \leftarrow \text{DDIM}(x_k, \widehat{x}_0, k)$
  if $M \neq \varnothing$ **then**
    $x_{k-1} \leftarrow S \odot \widetilde{x}_{k-1} + (1 - S) \odot q_{k-1}(m)$
  else
    $x_{k-1} \leftarrow \widetilde{x}_{k-1}$
  end if
end for
return $\textsc{Decode}(x_0)$

$k \in \{1, 2, 3\}$,

$$\text{R@}k = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}[\text{rank}(e_i \mid z_i) \leq k] \,. \tag{32}$$

We average 20 candidate samplings. This measures paired-caption alignment, not coverage of all valid motions for an ambiguous caption.

Multimodal distance (MM-Dist, lower is better). The mean paired distance $N^{-1} \sum_i \|z_i - e_i\|_2$ measures text–motion alignment, not within-caption diversity.

Fréchet Inception Distance (FID, lower is better). FID compares the full generated/GT embedding distributions, with means μ_g, μ_r and covariances Σ_g, Σ_r:

$$\text{FID} = \|\mu_g - \mu_r\|_2^2 + \text{tr}\Big(\Sigma_g + \Sigma_r - 2(\Sigma_g \Sigma_r)^{1/2}\Big) \,. \tag{33}$$

Features come from the G1 encoder, not Inception; FID is computed once per complete set, not per motion.

Diversity (reference diagnostic). We average $\|z_i - z_j\|_2$ over 300 distinct-motion pairs in each of 20 repetitions and compare with GT. Higher is not necessarily better. Multimodality, by contrast, requires multiple generations per fixed caption.

E.4.2 Kinematic Metrics

G1 FK is evaluated on the common predicted/GT prefix, without time warping. Within-clip scores are averaged equally across clips, then across repeated sampling runs. GT contacts are scoring masks only. These diagnostics measure adherence to reference contact timing as well as kinematic consistency: a valid generated swing can disagree with the GT support phase. They therefore do not isolate intrinsic foot slip under the generated motion's own contacts, or measure contact forces or controller success.

Foot Sliding (cm, lower is better). Ground is the fifth percentile of the framewise lower GT ankle height over the full target clip. Support requires height above ground at most 3.5 cm and planar speed at most 0.20 m/s; singleton contacts are removed. Sliding averages generated-ankle planar endpoint displacement over contact runs of both feet. It can miss within-run excursions and is not path length.

Contact Drift (cm, lower is better). Using four sole corners per foot, ground is the fifth percentile of pooled GT lowest-sole heights on the evaluated prefix. Support requires lowest-sole height within 3.5 cm above ground and sole-centroid planar speed at most 0.20 m/s, excluding singletons. Drift averages generated corners' planar endpoint displacements over corners, runs, and feet. Its geometry and mask differ from Sliding; neither measures maximum anchor excursion. Clips without valid runs contribute zero to both metrics.

Contact Height Error (cm, lower is better). For generated lowest-sole height $\hat{h}_{t,f}$ and the same sole ground g, average $|\hat{h}_{t,f} - g|$ over GT-contact frames and feet. Clips without selected contacts contribute zero.

Penetration (cm, lower is better). Average $\max(g - \hat{h}_{t,f}, 0)$ over *all* valid frames and feet. This measures depth, not the fraction of penetrating frames. All four errors are converted to centimeters. Figure 5 normalizes them separately by Heading-local means (Equation 27); no aggregate score is formed.

E.4.3 Editing and Human Ratings

Relative edit success. Let z_f^{src} and z_f^{edit} be the requested wrist's world-space FK height, $\sigma \in \{-1, 1\}$ the requested direction, and w_f the fixed sine-squared envelope. For the test request of 5 cm, success requires median signed gain $\mathrm{median}_{w_f \geq 0.8}\, \sigma(z_f^{\mathrm{edit}} - z_f^{\mathrm{src}}) \geq 3$ cm and height RMSE at most 2.5 cm over the complete editable interval, relative to the profile $\sigma\, 5 w_f$ cm. This measures the specified request, not token agreement or distance to a unique target pose. Hand height provides a directly measurable relative goal for this controlled protocol. It does not quantify the whole-body jump refinement shown qualitatively in Figure 1, nor does passing the height and preservation tests alone establish boundary smoothness or joint-limit validity.

Preservation and joint success. Scoring uses the private reference mask, never a method's predicted mask. Preservation requires at most 1% of protected joint/frame coordinates to change by more than 1°, protected FK position error at the 95th percentile at most 1 cm, maximum root displacement at most 1 cm and rotation change at most 1°, and unchanged duration. Joint success is the conjunction of edit and preservation success. All compared editors share projection; their protected joint MAE is zero. All requests and parse failures remain in the denominator. Same-cohort quality diagnostics are reported in Appendix D.5.

Automatic body–time mask overlap. Let $\widehat{S}_i$ and $S_i^\star$ be the predicted and private reference sets of editable frame–coordinate pairs within valid frames. Figure 6 reports coordinate-level macro IoU:

$$\mathrm{MaskIoU} = \frac{100}{N} \sum_i \frac{|\widehat{S}_i \cap S_i^\star|}{|\widehat{S}_i \cup S_i^\star|}. \tag{34}$$

An empty prediction against a nonempty reference scores zero. IoU is neither body-group accuracy nor motion-edit success.

All-turn editing success. Let $J_{i,r,t}$ indicate joint edit-and-preserve success at turn t of chain i in sampling replicate r. Each request is scored relative to the actual preceding output using the unchanged thresholds above. For $N = 597$ chains and $R = 3$ replicates, the curve in Figure 7 is

$$\mathrm{AllTurn}(d) = \frac{100}{NR} \sum_{i=1}^{N} \sum_{r=1}^{R} \prod_{t=1}^{d} J_{i,r,t}. \tag{35}$$

An earlier failure persists at greater depths; no failed chains are removed. Source-cluster uncertainty follows Appendix D.4.

Auxiliary whole-body and boundary diagnostics. World-space FK MPJPE averages all 29 actuated joint-body positions against the private IK witness without alignment, retaining coincident origins. The witness is not a unique valid edit. Boundary error measures edited-arm velocity change from source at both interval edges (20 FPS finite differences, deg/s). These diagnostics supplement request satisfaction.

User Study. The interface hides method names and identifies videos only by letters A–H; the letter-to-method mapping is stored separately from the survey page. For each prompt, method assignments to letters are randomly permuted once with a fixed seed. This order is shared across participants, rather than independently randomized for each participant. Videos appear together with the prompt in a grid, using a common G1 rendering setup, background, and camera configuration, without method-name overlays. Participants can pause and replay the clips and rate every video on two separate criteria: motion Quality and Text match.

The displayed three-point anchors are: **1 (poor)**, clearly unnatural or discontinuous motion with severe visible artifacts, or a mismatch with the main action or intent; **2 (fair)**, broadly plausible motion with local coordination defects or mild artifacts, or a broadly correct action with inaccurate details, direction, amplitude, or timing; and **3 (good)**, natural, continuous, coordinated motion with few visible artifacts, or accurate expression of the principal action, direction, and intent. The two criteria are scored independently; perceived motion quality is not a test of physical executability. Score definitions remain visible during rating, and all required ratings must be supplied to complete an item. Overall is the equal-weight mean of the two criteria. We first average each participant's ratings for a method, then report the mean and sample SD across participants.