EDITABLE ACTION WORDS FOR GENERATIVE MOTION EDITING

Anonymous authors
Paper under double-blind review

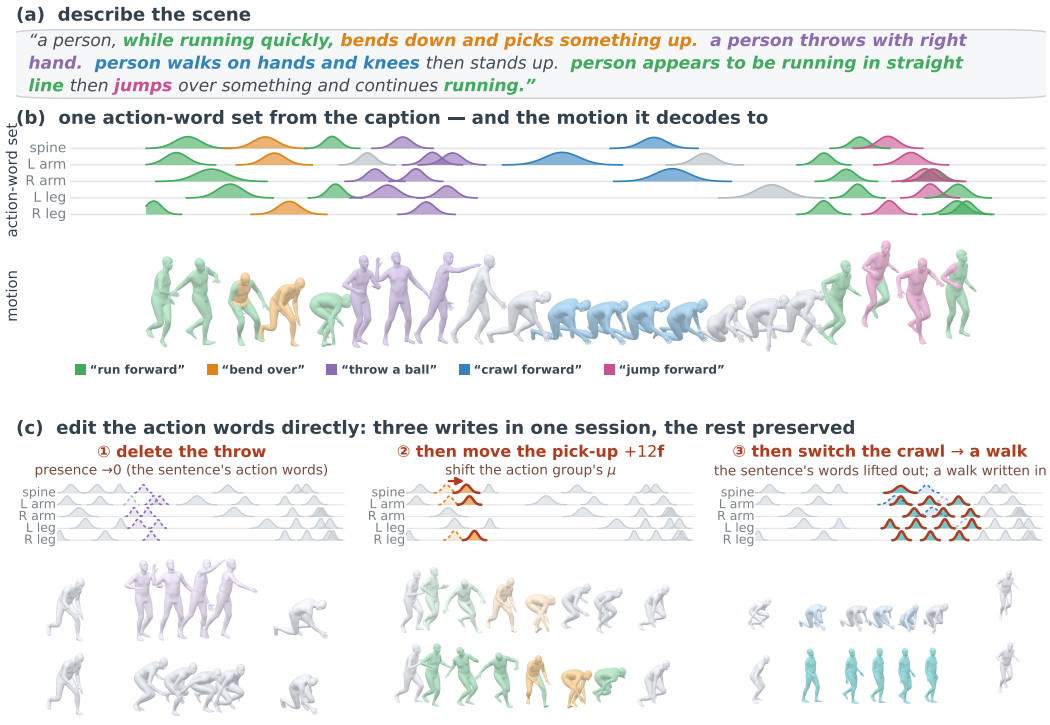


Figure 1: **An overview of the action-level generation and editing.** A caption becomes one *action-word set*: the colored timeline of named, addressable actions, decoded into the scene below it. Manipulating the action words revises the take in one session—*delete* the throw, *move* the pick-up later, *switch* the crawl into a walk—the contents outside each write preserved. **The supplementary video plays this session, and the benchmark, in motion.**

ABSTRACT

Text-based generative motion editing revises a generated clip from language, but mostly *within* an action, changing how it is performed. Editing the *set* of actions itself (adding, switching, moving, or removing one) has no direct handle: a generator returns a uniform grid of frames or tokens in which no action exists as an element, so the edit is approximated—a torn boundary, a disturbed rest, a sliding foot. We instead *generate the units editing needs*: a motion is a set of *action words*, one per action per body-part track. Each is a Gaussian on the clip's clock (a learned center, duration, and opacity) carrying a text-aligned content vector (*what*) and a manner residual (*how*); a named action is the group of action words that share it. The action is now a first-class element: the caption reads back as a legible timeline and an edit is a set operation that re-renders in distribution. Across *insert*, *switch*, *move*, and *delete* on an action-level editing benchmark, action words alone realize the operation while preserving the untouched motion and keeping foot contact physically plausible, where masked-infilling, attention, and instruction-tuned editors each give up one of these.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Text-to-motion models let a creator produce character motion from a description, such as "walk forward, jump, then wave" (Tevet et al., 2023; Chen et al., 2023; Guo et al., 2024; Zhu et al., 2023). Imagine previsualizing a shot through XR glasses, watching the motion play where the action will happen. The first clip is rarely the last: seeing the result, the creator wants to add a wave at the pause, switch the walk to a run, move the jump earlier, or remove a stray throw—while everything already approved stays put. These revisions target the *actions* themselves, not keyframes or frame ranges on a timeline. This paper makes a motion generator support them directly, on its own outputs and on captured clips alike.

Motion editors today revise a clip from a text instruction (Athanasiou et al., 2024; Jiang et al., 2025; Chen et al., 2024). But revising *one* action while keeping the rest is hard: nothing in an instruction picks out *this* throw in *this* clip, so instruction and attention editors apply the text by re-generating the clip—a new performance in its place, rather than that clip revised: afterwards the rest no longer reads as the actions it was. The edit can instead be bounded to a time window, where masked infilling locks everything outside (Tevet et al., 2023; Cohan et al., 2024). The window, however, becomes the user's problem: finding where the action starts and ends on a timeline that does not say. If the creator wants to change the throw, what they must specify is a start frame and an end frame—and even handed the window, a mask settles *when*, not *what* lands inside. We close this gap by making each nameable action an editable element of the generation itself.

Concretely, the generator produces the editable actions themselves (Fig. 1). A motion is a small set of *action words*, one per action per body-part track: a Gaussian on the clip's clock (*when*) carrying a text-aligned content vector (*what*) and a manner residual (*how*); a named action is a *group* of action words sharing content—the element an edit addresses. A caption (Fig. 1a) becomes one such set, the set reads back as a labeled timeline above the motion it decodes to (Fig. 1b), and an edit manipulates the words directly (Fig. 1c): deleting the throw drops its group, switching the crawl writes a walk's words in its place, and the motion re-renders from the set alone. Two learned maps produce the set—from a caption (a *prior*, T2W) or a raw clip (a *converter*, M2W)—so the unit is discovered, not annotated, and a captured clip edits the same way.

We evaluate on an *action-level editing benchmark*: four operations (insert, switch, move, delete) applied to a named action, scored by frozen judges for whether the edit *realizes*, whether the rest is *preserved*, and whether the result stays physically plausible, against the three editor families and a paste-and-regen reference. Action words are the only object holding every side of the frontier at once—realize, preserve, seam, contact: infillers preserve but miss the target, attention and instruction editors realize only by rewriting the clip, and paste-and-regen keeps the rest by construction but tears what it regenerates—or, smoothed, slides planted feet.

Our contributions are **(1)** the action word, a set substrate whose element *is* the action a caption names, with a converter (M2W) that constructs the units from raw motion (§3.1); **(2)** a native generator over it, trained from scratch on complete pairs with identity owned by the action words (§3); and **(3)** an action-level editing benchmark and evaluation under frozen judges, on which action words realize the operations in distribution while preserving the untouched motion (§4). The problem was not generating the action; it was that the operations that edit it had no element to act on. An action word is that element.

2 EDITABLE ACTION UNIT

An action-level edit can be approximated on certain objects, but what the object provides determines what the edit costs. We state four *design goals*, then survey prior work by the *object* it edits.

2.1 DESIGN GOALS

What should the object provide for an action-level edit? Each goal an object lacks is paid for somewhere—the boundary, the untouched rest, or the target itself. Two are the crux. **(DG1) Spatiotemporal locality**: the unit holds a definite extent in time and on the body, so an edit touches that support alone rather than the whole clip. **(DG2) Action correspondence**: the unit is a single named action with a text-aligned identity, so an edit can name *which* action; manner-only supervision does

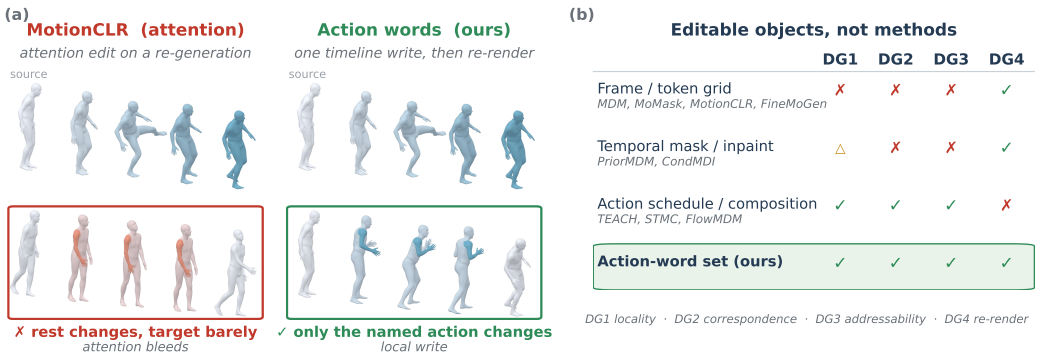


Figure 2: **Editable objects, not methods.** *(a)* One edit (switch the kick to a throw): MotionCLR (Chen et al., 2024) edits attention on a re-generated clip, so the rest changes while the target barely does (§4.2). *(b)* Each editor's *object* against the design goals DG1–DG4: only the action-word set holds all four.

not isolate this. Editing then adds two more. **(DG3) Addressability**: the units are discrete, individually writable handles read back as a labeled timeline, so an edit is a write to a named element rather than a re-prompt. **(DG4) Re-renderability**: editing the set and decoding gives one seamless motion from the set alone, with no boundary to splice. *Local persistence*—one write leaving the rest untouched—follows from DG1 with DG4. DG1 and DG3 make the units *direct-manipulation* handles (Shneiderman, 1983), and two-way production (read off a clip, or sampled from a caption) is *analysis-by-synthesis* (Yuille & Kersten, 2006). No editable object in use meets all four (Fig. 2).

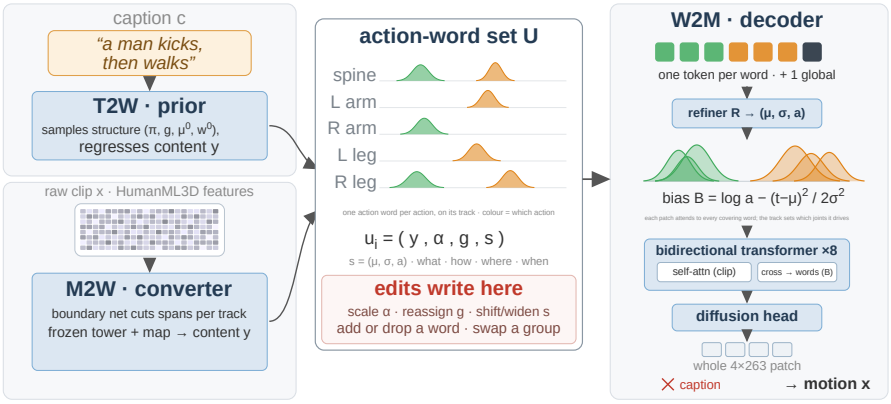
2.2 Prior editable objects

Motion editing. Most text-based generative motion editing is *intra-unit*—it changes how an action is performed, not which actions occur—and unit-level attempts work around the missing action element. Positional inpainting and keyframe in-betweening (Shafir et al., 2024; Cohan et al., 2024) take a mask: timing without semantics (against DG2). Attention steering (Chen et al., 2024) edits a transient control inside a fresh generation, not an addressable element (against DG1, DG3); caption re-generation and timeline composition (Petrovich et al., 2024; Barquero et al., 2024) re-roll or stitch whole segments, so approved motion does not persist (DG1) and the boundary is smoothed rather than rendered (DG4). Instruction editors, including LLM rewrites of pose codes (Athanasiou et al., 2024; Goel et al., 2024; Li et al., 2025b; Guo et al., 2025; Jiang et al., 2025; Wu et al., 2025a; Huang et al., 2024; Wu et al., 2025b), learn paired how-tweaks over the whole clip with no handle on *which* action—an interval hint (Jiang et al., 2025) selects where to restyle, not which action to add or switch; noise-space optimization (Karunratanakul et al., 2024) relocates under an objective, not a name. None preserves a source clip while realizing an action-level change (§4.2).

Generation substrates. The discrete line quantizes fixed windows (Zhang et al., 2023a; Guo et al., 2024; Pinyoanuntapong et al., 2024; Jiang et al., 2023; Zou et al., 2024; Meng et al., 2025; Yan et al., 2026); diffusion (Tevet et al., 2023; Zhang et al., 2024; Chen et al., 2023) adds semantics as conditioning (Jin et al., 2023; Nazarenus et al., 2026; Hong et al., 2025). FineMoGen (Zhang et al., 2023b) comes closest, with body-part temporal Gaussians—but as fixed-width templates inside the denoiser, not an object a person writes to (against DG3), and an edit still resamples the clip (against DG1, DG4). Latent *motion primitives* (Zhao et al., 2025) and neural fields (He et al., 2022) name a fixed chunk of clock, not an action. Ours borrows 3D Gaussian splatting (Kerbl et al., 2023; Wu et al., 2024), pruned by need, on the time axis; no prior work makes this variable-rate, action-aligned boundary its generated object.

Units as interfaces or labels. Timeline systems generate from interval structure a person supplies (Athanasiou et al., 2022; Petrovich et al., 2024; Barquero et al., 2024): the boundary is not a variable and the segment carries a label, not content. Per-frame text or codes (Li et al., 2025a; Huang et al., 2024; Liu et al., 2024) are fixed-rate or closed-inventory; variable-length latent segments date to CompILE (Kipf et al., 2019); robotics movement primitives put Gaussian bases on a trajectory's clock (Paraschos et al., 2013), per-DOF control with no semantic identity. We adopt speech's teacher-re-estimation recipe (Hsu et al., 2021; Baade et al., 2024) but diverge on the object,

Figure 3: **Architecture.** Two producers, one object: the prior (T2W) samples the action-word set from a caption, the converter (M2W) reads it off a raw clip, and editing writes to the set. The decoder (W2M) renders motion from the set alone (DG4), each action word a key/value token whose learned (μ, σ, a) biases cross-attention over 49 patch tokens; no caption channel exists.



borrowing *intrinsic* unit evaluation from language and speech (Mikolov et al., 2013; Nguyen et al., 2020; Zhu et al., 2024; Ramanujan et al., 2024).

3 METHOD

The design goals are met by one object and three maps around it. The *action-word set* (§3.1) carries the localized, named, separately writable coordinates DG1–DG3 ask for; a *converter* (M2W) reads a set off a raw clip and a *prior* (T2W) samples one from a caption, so the object is produced either way; a *decoder* (W2M) renders motion from the set alone, trained on complete pairs, so the set owns and re-renders the motion (DG4). Either producer yields a set, a person rewrites it, and W2M renders whatever it is handed (§3.5).

3.1 THE ACTION WORD

The design goals (§2.1) fix the object. A T-frame clip is a set of *action words*,

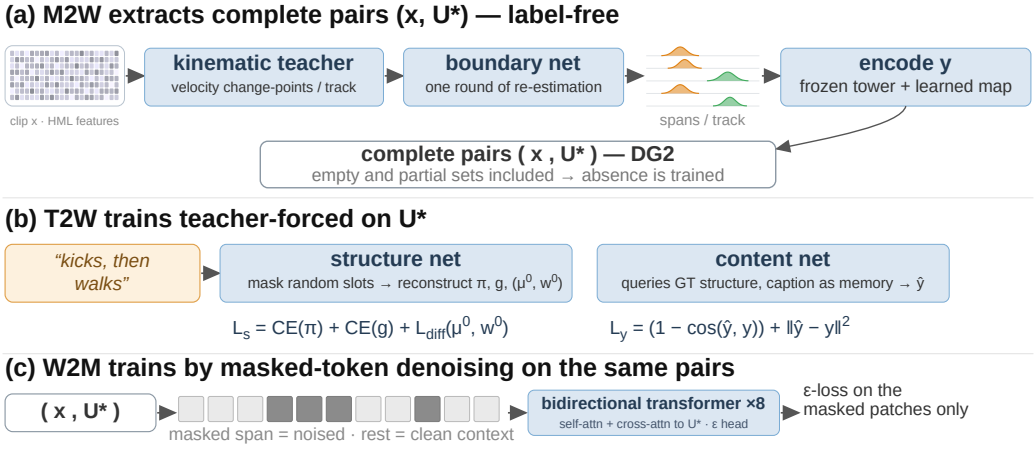
$$x \approx \text{Decode}\big(\{u_1, \ldots, u_K\}\big), \qquad u_i = \big(\underbrace{y_i}_{\text{what}}, \underbrace{\alpha_i}_{\text{how}}, \underbrace{g_i}_{\text{where}}, \underbrace{s_i}_{\text{when}}\big), \tag{1}$$

where y_i is a continuous content vector, α_i a manner residual, g_i a body-part track, and $s_i = (\mu_i, \sigma_i, a_i)$ a temporal Gaussian $a_i \exp(-(t-\mu_i)^2/2\sigma_i^2)$. Content lives in a jointly aligned text–motion retrieval space (Petrovich et al., 2023) and the timeline's labels are this vector read back through text (DG2). A presence bit makes the set variable-size. An action word occupies one track, so a named action moving several body parts is a *group* of action words sharing content—the element an edit addresses, its members switching, retiming, or vanishing together. Fig. 9 (Appendix H) contrasts the object with the frame and token grids.

3.2 THE CONVERTER (M2W): CONSTRUCTING THE EDITING UNIT

The three maps chain rather than train jointly (Fig. 4). M2W runs first and trains *once*; T2W and W2M then train from scratch on *the same* extracted sets. M2W is where the editable unit is *constructed*, and it upper-bounds the system by fixing the boundaries, the content, and where nothing acts. Boundaries come from motion alone: a kinematic *teacher* marks change-points of per-track joint velocity, dense and label-free, and distills a per-frame *boundary network*, re-estimated once on its own proposals (Hsu et al., 2021) to sharpen the cuts. Peaks cut variable-length spans per track, one per action on the track it moves—a whole-body action appears as co-spanning spans (the *group* of §3.1)—and a presence bit marks tracks where nothing acts, an explicit, trained condition (Appendix A).

Figure 4: **Training.** **(a)** M2W extracts an action-word set from every clip, label-free, giving the complete pairs (x, U^*). **(b)** T2W trains its structure and content networks teacher-forced on those sets. **(c)** W2M trains by masked-token denoising on the same pairs.



Each span is then *named*: its content is the span's motion pushed through the frozen retrieval tower ϕ and one learned projection P_ψ,

$$y_i \;=\; P_\psi\,\phi\big(x[\,t_i^s{:}t_i^e\,]\big), \tag{2}$$

where a multiple-instance objective aligns spans to caption *fragments* and trains *only* P_ψ—tower, boundaries, and spans untouched—so captions name the units without training the segmenter. The coarse window (μ^0, w^0) is read off the span with no network, and the decoder later refines it into the rendered Gaussian (§3.4). Running M2W over the corpus yields the complete pairs (x, U^*)—full, partial, and empty sets alike—that DG4 requires (Fig. 7, Appendix A); at edit time the same steps run as one forward pass on an unseen clip, no per-clip optimization, no text (§3.5).

3.3 The prior (T2W): sampling a set from text

T2W models $p(U \mid c)$ over $M{=}24$ slot positions—this sampler's ceiling only (corpus P99; the representation carries no length limit), presence bits keeping the set variable-size. The caption c enters as token features and a pooled embedding of the same retrieval text tower that defines the content space (DG2). T2W factorizes where the ambiguity is, sampling *structure* and regressing *content*:

$$p_\theta(U \mid c) \;=\; \underbrace{p_{\theta_s}\big(\{\pi_i, g_i, \mu_i^0, w_i^0\}_{i=1}^{M} \,\big|\, c\big)}_{\text{masked set transformer (structure)}} \cdot \underbrace{\textstyle\prod_{i:\,\pi_i=1} \delta\big(y_i - f_{\theta_y}(i, \{\pi, g, \mu^0, w^0\}, c)\big)}_{\text{content regressor}}, \tag{3}$$

where structure is presence π_i, track g_i, and a coarse window (μ_i^0, w_i^0) that the decoder refines into s_i (§3.4), and δ is the Dirac mass, so once structure is sampled content is deterministic. Eq. 3 is an empirical choice, not a loss (Appendix E): sampling content pays a distribution cost, while regressing structure discards real ambiguity. Both networks teacher-force on M2W's structure so errors do not compound, the structure net reconstructing randomly masked slots (windows by small per-slot diffusion heads (Meng et al., 2025), so placement is sampled, not averaged) and the content net regressing each present slot's $\hat{y}_i$ (Appendix B). At sampling, structure fills over a few confidence-ordered passes, content in one regressor pass; the regressor predicts the conditional *mean* of y, so sampled content can land off-manifold—an inference-time snap corrects it (§3.4).

3.4 The decoder (W2M): rendering motion from the set

W2M models $p_\phi(x \mid U)$ conditioned on the action-word set alone (DG4); frames no action word covers are filled from its own prior. It is a masked generative transformer in the MaskGIT family (Chang et al., 2022) over *continuous* tokens (Meng et al., 2025): the clip becomes patch tokens of

four frames (a linear map, no codebook), read by a bidirectional transformer with cross-attention to the action words, a per-token diffusion head denoising each patch. Cross-attention from a patch token at frame center t to action word i is *splatted*,

$$B_{t,i} \;=\; \log a_i \;-\; \frac{(t-\mu_i)^2}{2\sigma_i^2}, \tag{4}$$

so a token draws on the action words that cover it in proportion to their opacity, with one learned global token as the only other key. The temporal parameters come from a small *refiner* $(\mu_i, \sigma_i, a_i) = R_\theta(y_i,\, g_i,\, \mu_i^0,\, w_i^0)$, learned end-to-end through the splatted logit with no separate geometry objective—point-based graphics' differentiable-rendering move on the time axis—so the s_i of Eq. 1 is learned, not rule-defined. *Training* is masked-token denoising on the complete pairs under a mask schedule spanning edits, generation, and the CFG anchor; three in-model additions replace any post-hoc smoothing (Appendix C). Sampling decodes the set in one guided pass; a caption-sampled set adds a train-free content-snap and a half-patch-offset fused decode (Bar-Tal et al., 2023) closing its residual gap (Appendices B, D).

3.5 Editing is manipulating the set

Because W2M's condition is the action-word set itself, editing is rearranging that set and re-rendering, with no inversion or caption rewrite (DG3, DG4). Each operation is one move on the tuple $u_i = (y_i, \alpha_i, g_i, s_i)$ or on set membership: scale manner α, reassign track g with a mirror map, retime or resize the window s, or change membership (insert, delete, compose, and *switch*, which replaces the target's words with a donor group—the new action's own words). Two rules keep the moves well-posed. First, whole-body edits act on the *action group*—the target plus its span-overlapping words—content-gated per op: switch and delete take only the target's *same-content* members, while move shifts the co-spanning slots as one beat. Second, content is portable across clips because every vector shares one space. An edit re-decodes the clip from the edited set—no frames are frozen; the untouched action words re-render their own motion, so the rest is kept as *actions*.

4 Experiments

4.1 Setup and baselines

Editing baselines are the published families without an action element: the masked infillers MDM inpaint (Tevet et al., 2023) and CondMDI (Cohan et al., 2024), the attention editor MotionCLR (Chen et al., 2024), and the instruction-tuned editors MotionLab (Guo et al., 2025), MotionReFit (Jiang et al., 2025), and SimMotionEdit (Li et al., 2025b). As a reference for the workaround we construct *paste-and-regen*. Outside the edit window the source is copied; the window—the target group's span, the same element ours edits—is filled per operation. *Insert* and *switch* generate the target motion at window length from the donor phrase alone (MDM (Tevet et al., 2023)) and paste it, full body. *Delete* inpaints the window away: the same sampler run text-free, in-betweening the locked surround. *Move* copies the window's frames Δ later, then inpaints the vacated window the same way. A smoothed variant applies the per-op seam-minimizing smoother (Appendix K). Every window-consuming baseline—paste-and-regen, both infillers, and MotionReFit's given-window variant—receives this same window, and judging crops remain the target span for every method. All methods inherit the identical clips, windows, and donor specification—in each method's native form: the donor phrase for the caption- and instruction-driven baselines, the donor's own action words for ours; per-operation mechanics, ours included, are documented in Appendix K, and generation quality against text-to-motion baselines in Appendix E.

4.2 An action-level editing benchmark

Standard suites score only generation, so we build one for action-level editing: it measures how well each representation *expresses* the operations. On 140 HumanML3D clips (insert: 88) it applies the four operations to a named action, scored by a frozen TMR judge (Petrovich et al., 2023). Insert targets audited-empty gaps, adding an action rather than overwriting one (§H.1). Four metrics ask one editing question each. *Realize*↑ (did the change take on the named action?): the fraction of

Table 1: **The action-level editing benchmark** (n=140; insert 88), grouped by the object each family edits. Metrics (*real.*ize, *sem.* survival, seam, skate) are defined in §4.2; a blank cell (—) means that operation is *not expressible* in that method. The last row edits caption-generated sets.

	Insert				Switch				Move				Delete			
Method	real.↑	sem.↑	seam↓	skate↓	real.↑	sem.↑	seam↓	skate↓	real.↑	sem.↑	seam↓	skate↓	real.↑	sem.↑	seam↓	skate↓
Instruction-tuned (the edit is a text revision from paired examples)																
MotionLab	0.15	1.00	2.9	.065	0.49	0.43	2.8	.026	0.19	0.99	3.3	.053	0.60	0.43	2.9	.028
MotionReFit	0.50	0.10	9.0	.551	0.41	0.09	8.8	.421	—	—	—	—	0.89	0.09	8.7	.406
given the target window	0.01	1.00	47.2	.066	0.41	0.91	28.2	.089	—	—	—	—	0.89	0.91	25.3	.087
SimMotionEdit	0.43	0.02	1.7	.061	0.52	0.00	1.1	.052	—	—	—	—	0.94	0.01	1.1	.057
Attention manipulation (the edit steers a re-generation)																
MotionCLR	0.60	0.18	2.7	.147	0.71	0.05	2.5	.060	0.52	0.16	5.6	.189	0.69	0.33	2.8	.068
Masked infilling (the edit is a masked window)																
MDM inpaint	0.57	1.00	9.3	.072	0.58	0.95	9.4	.089	0.23	0.98	4.1	.080	0.86	0.93	7.4	.073
CondMDI	0.23	1.00	6.1	.086	0.54	0.95	5.4	.116	—	—	—	—	0.60	0.95	5.7	.117
Paste-and-regen (a constructed reference; mechanics in §4.1)																
Paste-and-regen	0.86	1.00	41.2	.084	0.86	0.90	33.1	.081	0.86	0.96	9.7	.078	0.83	0.92	7.2	.073
+ smoothing	0.81	0.99	0.1	.111	0.86	0.89	1.8	.124	0.87	0.93	1.1	.079	0.84	0.90	1.1	.071
Action-word set (the edit is a set write, re-rendered)																
Action words (ours)	0.96	1.00	2.9	.064	0.97	0.96	1.6	.049	0.76	1.00	1.1	.035	0.94	0.97	0.3	.025
on generated (the prior) sets	0.94	1.00	2.9	.033	0.92	0.86	1.3	.037	0.75	0.98	1.1	.031	0.81	0.95	0.3	.030

clips the op succeeds on under the judge (insert and switch: the donor becomes retrievable on the window's crop, a TMR donor-lift; move: presence at the new window minus at the old, so residue left behind lowers it; delete: adherence to the removed action drops on the target span). *Sem.*↑ (does the rest still *read* as its actions?): the fraction of action words outside the edit zone whose own-phrase reading survives the edit (paired frozen-TMR read, broken if it drops by more than 0.10; only words the source reads at ≥ 0.30 count; the geometric companion—joint error in cm—is Table 5). *Seam*↓: how abrupt the join at the edit boundary is—the peak boundary jerk in a $\pm$3-frame band minus the clip's own interior baseline, in cm (Barquero et al., 2024)—read against the band real action entries occupy, so a torn boundary (above it) and a lifeless over-smoothed one (below it) both count against (Appendix K). *Skate*↓: the foot-skate ratio (source-re-decode floor .033; raw capture reads .055 under the same detector, Table 9).

4.3 EDITING PERFORMANCE

Because the decoder's condition *is* the action-word set, an edit is an in-distribution set operation, and action word edits (Table 1) hold the full frontier (realize the edit, keep the rest reading as its actions, keep the boundary in the natural band, keep feet planted).

The learned editors we could run each miss the edit for a family-specific reason. Infillers preserve, and realize exactly as much of the edit as the construct itself carries: emptying the window *is* a delete (0.60–0.86), the locked surround plus the caption cues a switch to about half (0.54–0.58), and insert peaks at 0.57 even with the caption reduced to the donor phrase alone (0.23 for CondMDI)—a mask fixes *when*, but *what* arrives only as often as text delivers it, and move regenerates the shifted window *without vacating the original* (0.23; CondMDI's one window cannot express a relocation). For the attention and instruction editors, the "edit" is a fresh performance in the clip's place: re-read afterwards, most of the untouched action words no longer read as their own action (*sem.* 0.00–0.43), MotionLab barely responds to the instruction (Appendix K), and the 8–56 cm joint error (Table 5) is the geometric footprint of that rewrite, not its cause. Even handed the target span, MotionReFit ($[t_s, t_e]$, rest as source) only shifts the bind: the outside stays source-verbatim and the emptied window does delete (0.89), but the span never integrates (25–47 cm seam) and nothing new lands in it (insert 0.01, switch 0.41).

Paste-and-regen bounds the workaround itself. Handed our windows and donor phrases (§4.1), the splice realizes every operation at 0.83–0.87, with move (0.86–0.87, verbatim source frames re-pasted at the new time) the one operation above ours (0.76; vacating matches—the paste's GT frames simply out-read a re-rendered instance). What it cannot buy is the rest of the frontier. On insert and switch its generated material is bound by the sampler itself—86% of the shared windows are shorter than the generator's 2-second training minimum, where direct sampling collapses and the deployed resample trick recovers the phrase at a physics cost (Table 10)—so the rates stop



Figure 5: **Table 1, rendered: one case per operation**; each edited cell reports *edit* ✓/×, *sem* (this case's surrounding action words still reading as themselves, as in Table 1), *seam* (cm). Given our group windows, donor phrases, and its text-free fill (§4.1), paste-and-regen realizes all four cells shown here; what the benchmark prices is elsewhere—raw, its pastes tear (Table 1: seam 33–41 cm), and this smoothed variant slides feet and stretches bones (skate .111/.124, bone 0.71/1.06). MotionCLR realizes only by rewriting the clip (sem 0.00–0.67 here); ours realizes each edit and keeps the rest reading as it did.

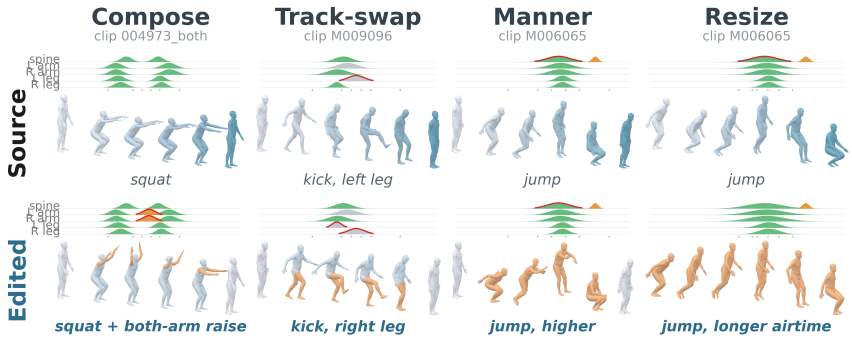


Figure 6: **Composition and attribute edits, qualitatively** (source vs. ours; edited action saturated, timelines mark the touched action word red). *Compose* overlays a both-arm raise over a squat, *track swap* mirrors the kick, *manner* raises the jump apex, and *resize* (the same clip as manner) widens the jump group: still airborne where the source lands.

at 0.86 vs. ours 0.96/0.97, and delete separates the same way (0.83 vs. 0.94; ours removes *every* instance where the fill empties one window, Fig. 5). The pastes also tear the boundary (33–41 cm seam, Appendix K)—and the tears sit next to the neighbors' readings (*sem.* 0.89–0.96 vs. ours 0.96–1.00)—and the text-free fills carry the sampler's own physics (move/delete bone jitter 0.6–0.7 vs. ours 0.19–0.28); smoothed to its per-op best, the seam collapses (33.1 → 1.8) while foot-skate rises to the table's worst on the generated ops (.111/.124 vs. ours .025–.064) with the largest insert/switch bone jitter (Table 9). Each cell the workaround wins is bought at a boundary or plausibility cost the frontier prices.

Ours, beyond the cells. All four edits act on the action group (§5), delete removes every same-content instance (0.77 → 0.94), content is portable across clips (cross-decode FID 0.044), and caption-generated sets edit within 0.14 of the extracted-set realize (Appendix E).

4.4 Qualitative results, sessions on captured clips, and ablations

Sessions on captured clips. Because the converter reads the set off any raw clip, the same writes run on *captured* motion—no caption anywhere, the clip re-decoded from the edited set in place. Fig. 12 shows a cleanup pass: a boxing-form take mixes kicks into the punching, and one *delete* removes them, whole group, the decoder bridging the vacated window from the remaining set. A timing pass pulls a mid-take crossing -12 frames, its beat traveling whole (Fig. 13); and variant iteration on a karate combo *repeats* the kick group a beat later (double kick $\rightarrow$ triple), then *switches* the first kick into a punch group (Fig. 14)—takes accumulate on one timeline rather than through re-capture (Appendix I).

Beyond the four operations, action word edits compose across tracks, swap a body-part track, and tune manner as set writes (Fig. 6; Appendix H.1); the content coordinate is open—walk$\rightarrow$run interpolation morphs the gait and the bank-level offset $\bar{y}_{\mathrm{run}} - \bar{y}_{\mathrm{walk}}$ flips the read-back with no donor (Fig. 8, Appendix H.2)—and the learned centers land within 1–2 frames of human impact annotations (Jung & Lee, 2026) (Appendix F). Ablations (opacity budget, one vs. five tracks, background, decode head) are in Table 2 (Appendix C). Generation holds up on the standard suite: a clip decoded from its own action words reaches FID 0.040 at R@3 0.798 (GT 0.806), below every measured baseline; the caption-sampled pipeline (T2W$\rightarrow$W2M) reaches R@3 0.843 at FID 0.092—ahead of MoMask and SALAD on retrieval, behind them on FID (Appendix E).

5 Discussion

Three boundaries mark where the current instantiation can be improved (expanded, with qualitative evidence, in Appendix J). The action-group rule is fixed—editing one action word carries its co-spanning neighbors (single-word relocation drops move $0.76 \rightarrow 0.30$, Fig. 10)—the right default for whole-body actions; the cost, per-part freedom within an action, a learned finer unit would free. Attribute edits are graded: *resize* extends ballistic actions while sustained ones barely lengthen (duration lives in content, not the window alone), and *compose* keeps both actions alive but is noisiest on same-track overlaps (§H.1). And the minutes-long scene is ahead of the evidence (HumanML3D caps clips at ten seconds): scaling needs hierarchical grouping, a streaming decoder, and interaction design at the timeline.

Two directions stand out. *Geometry*: content is tuned to *what* action occurs—not to fine spatial detail (how far a kick travels, where a hand lands), to *traversal* (distance and heading are inferred by the decoder, not exposed as a coordinate an edit could dial, and long re-renders can drift or turn between correct actions), or to *dynamics* (a strongly resized jump stretches its flight kinematically rather than re-solving ballistics, reading as reduced gravity). A traversal coordinate on the word (or a root-path channel beside the set) is the natural extension, and a geometry- and physics-aware content space and judge would sharpen both. *Decomposability*: co-spanning action words share content across tracks, so a whole-body action edits as one group; track-level writes on one-track actions work today (track swap, per-track compose, Fig. 6), yet factorizing the tuple into a content core plus per-track residuals would isolate per-part edits and clean up composition.

6 Conclusion

In sum, *action words* make the editable unit the action itself: a clip is a variable-size set of temporal Gaussians—one action's content, manner, track, and learned timing each—that a native generator renders from the set alone. A caption becomes a legible timeline of directly manipulable action words, whose one-write edits realize *insert*, *switch*, *move*, and *delete* at 0.76–0.97, preserving the rest and keeping contact—a frontier no baseline holds jointly—while a clip re-decoded from its own words reaches FID 0.040 (caption sampling: R@3 0.843 at FID 0.092). Making the action the object, not a better generator, turns an action-level edit into a write.

ETHICS STATEMENT

This work uses only the public HumanML3D benchmark (AMASS- and HumanAct12-derived captioned motion capture); no new human data was collected, and the clips carry no personally identifying information.

AI USE STATEMENT

We used generative AI tools to assist in carrying out experiments (authoring and running evaluation scripts under our direction) and to build visual analysis tools used to inspect results, as well as for literature search and editing of author-written text. All claims, numbers, and artifacts were reviewed and verified by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The representation, the segmenter, the generator, and the evaluation protocol are described in Sections 2–3, with architectures, training recipes, and metric definitions in Appendices A–E. The corpus measurements are reproduced by single scripts over the public HumanML3D captions. Code, segmenter checkpoints, pre-registered operation sets, and the evaluation code will be released.

REFERENCES

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *3DV*, 2022.

Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J. Black, and Gül Varol. MotionFix: Text-driven 3d human motion editing. In *SIGGRAPH Asia*, 2024.

Alan Baade, Puyuan Peng, and David Harwath. Syllablelm: Learning coarse semantic units for speech language models. *arXiv preprint arXiv:2410.04029*, 2024.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023.

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings (flowmdm). In *CVPR*, 2024. arXiv:2402.15509.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022.

Ling-Hao Chen, Wenxun Dai, Xuan Ju, Shunlin Lu, and Lei Zhang. MotionCLR: Motion generation and training-free editing via understanding attention mechanisms. *arXiv preprint arXiv:2410.18977*, 2024.

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

Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.

Purvi Goel, Kuan-Chieh Wang, C. Karen Liu, and Kayvon Fatahalian. Iterative motion editing with natural language. In *SIGGRAPH*, 2024.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. *arXiv preprint*, 2022.

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

Ziyan Guo, Zeyu Hu, Na Zhao, and De Wen Soh. MotionLab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *ICCV*, 2025.

Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *NeurIPS*, 2022.

Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *CVPR*, 2025. arXiv:2503.13836.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *ECCV*, 2024.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Dynamic motion blending for versatile motion editing. In *CVPR*, 2025.

Peng Jin, Yang Wu, Yanbo Fan, Zhongqian Sun, Yang Wei, and Li Yuan. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. In *NeurIPS*, 2023.

Euijun Jung and Youngki Lee. Per-stroke temporal control for text-to-motion via action units and action-detection guidance. *arXiv preprint arXiv:2607.15717*, 2026.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *CVPR*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Thomas Kipf, Yujia Li, Hanjun Dai, Vinicius Zambaldi, Alvaro Sanchez-Gonzalez, Edward Grefenstette, Pushmeet Kohli, and Peter Battaglia. Compile: Compositional imitation learning and execution. In *ICML*, 2019.

Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. In *International Conference on 3D Vision (3DV)*, 2025a.

Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. SimMotionEdit: Text-based human motion editing with motion similarity prediction. In *CVPR*, 2025b.

Xinpeng Liu, Yong-Lu Li, Ailing Zeng, Zeyi Zhou, Yang You, and Cewu Lu. Bridging the gap between human motion and action semantics via kinematic phrases. In *ECCV*, 2024.

Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. In *CVPR*, 2025.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *NAACL*, 2013.

Eric Nazarenus, Chuqiao Li, Yannan He, Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Actionplan: Future-aware streaming motion synthesis via frame-level action planning. *arXiv preprint arXiv:2603.13500*, 2026.

Tu Anh Nguyen et al. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv preprint arXiv:2011.11588*, 2020.

Alexandros Paraschos, Christian Daniel, Jan Peters, and Gerhard Neumann. Probabilistic movement primitives. In *Advances in Neural Information Processing Systems*, 2013.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023.

Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPR Workshop on Human Motion Generation (HuMoGen)*, 2024. arXiv:2401.08559.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. MMM: Generative masked motion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, 2021.

Vivek Ramanujan et al. When worse is better: Navigating the compression-generation tradeoff in visual tokenization. *arXiv preprint arXiv:2412.16326*, 2024.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior (priormdm / doubletake). In *ICLR*, 2024. arXiv:2303.01418; DoubleTake inference-time handshake stitching.

Ben Shneiderman. Direct manipulation: A step beyond programming languages. *IEEE Computer*, 16(8):57–69, 1983.

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

Bizhu Wu, Jinheng Wang, Lingxiao Zhang, Bingbing Ni, Xin Liu, Xiaodong Yang, and Ke Zhang. MG-MotionLLM: A unified framework for motion comprehension and generation across multiple granularities. In *CVPR*, 2025a.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024.

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with LLMs. In *ICLR*, 2025b. arXiv:2405.17013.

Sheng Yan, Yong Wang, Xin Du, Junsong Yuan, and Mengyuan Liu. Language-guided transformer tokenizer for human motion generation. *arXiv preprint arXiv:2602.08337*, 2026.

Alan Yuille and Daniel Kersten. Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. *arXiv preprint*, 2023a.

Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. In *NeurIPS*, 2023b. arXiv:2312.15004.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint*, 2024.

Kaifeng Zhao, Gen Li, and Siyu Tang. Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *ICLR*, 2025. arXiv:2410.05260.

Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99%. *arXiv preprint arXiv:2406.11837*, 2024.

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. arXiv:2307.10894.

Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. In *ECCV*, 2024.

A Segmenter and Content Space

Construction. The segmenter is teacher-first and label-free throughout. A *kinematic change-point teacher* proposes variable-length spans per body-part track from each track's kinematics, with no action inventory and no captions; from it we distill a *per-frame boundary network* (a dense per-frame, per-track boundary head), then run one re-estimation pass: the head re-reads each proposal's onset and end from its own dense start/end fields, one round and no second training stage (the spirit of the speech pseudo-label recipe (Hsu et al., 2021)). Each span initializes its temporal geometry at the span middle before the refiner (§3.4) takes over. Training uses the HumanML3D train split only; human segments (Punnakkal et al., 2021) referee boundary quality without supervising it (Fig. 7). The distilled head reaches recall 0.794 at IoU 0.3 with 3.52 proposals/track (teacher 0.807–0.849).

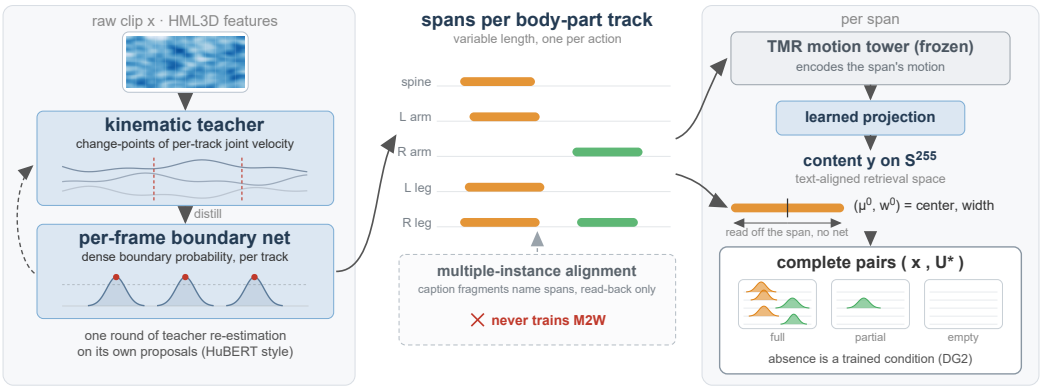


Figure 7: **M2W, raw clips to complete pairs, label-free.** A kinematic change-point teacher proposes per-track spans and distills a per-frame boundary network, re-estimated once on its own proposals (the speech pseudo-label recipe); peaks cut variable-length spans per body-part track, one per action. Each span is encoded by the frozen retrieval tower and a learned projection into its content y, with the coarse window (μ^0, w^0) read off the span with no network; caption fragments name spans through a multiple-instance alignment that trains *only* the naming projection P_ψ—boundaries and spans receive no caption supervision. The result is the complete pairs (x, U^*)—full, partial, and empty sets alike—so absence is a trained condition (DG2).

Content space. A span's *content* y is a vector in the text-aligned retrieval space, its motion encoded by the frozen retrieval tower and a learned projection, so alike spans land near each other and text maps in. The *manner residual* α is the component orthogonal to the content direction (the *how* apart from the *what*). Caption fragments enter through a multiple-instance InfoNCE, the positive being the best in-clip fragment–span match (both directions), and it trains *only* the two projections and a temperature: the towers, the boundary net, and the spans stay frozen, so text shapes the content coordinates, not the segmentation, and read-back naming is a post-training argmax. The learned projection reaches 0.558 head-verb R@1 vs. 0.374 zero-shot, grounding 0.89; a VQ code (0.222) and per-span queries (0.179) were worse, and the residual separates variants at d=1.83 split-half.

B Architecture and Training Details

Field dimensions. Widths, heads, and losses are in §3; here we record the dimensions those leave implicit. Content $y_i \in \mathbb{R}^{256}$ takes the native width of the retrieval tower that defines the space, so no projection width is introduced, and the manner residual $\alpha_i \in \mathbb{R}^{16}$ sits at the smallest width whose reconstruction of held-out manner variation stopped improving (8 underfits limping and stiff

variants; 32 does not improve them). The element is one-dimensional in time: it plays the role of a splat in point-based rendering, position, scale, and opacity, contributions superposing and pruning, on the clock rather than in space. The refiner R_θ (§3.4) is a 3-layer MLP (290→256→256→3, GELU) over the concatenation of content y_i, a 32-d track embedding, and the normalized coarse window (μ_i^0, w_i^0); its last layer is zero-initialized, so it emits the seed Gaussian at init and departs from it only as the decoding loss pulls it, its three outputs mapping to center $\mu_i = \mu_i^0 + \tanh(\cdot)\, w_i^0/2$, scale $\sigma_i \in [1.5\,\mathrm{f},\, w_i^0]$, and opacity logit a_i. The decoder takes nothing besides the set: frames no action word covers are filled from its own prior through the learned global token. The patch (not the frame) is the token because one shared head (a small MLP run at every masked patch) then emits four coherent frames at once, removing frame-rate jitter, and attention over the clip costs $16\times$ less; the masked form makes the same pass iterable, any known patches serving as clean context for the rest. Clips shorter than the 196-frame canvas are zero-padded under a validity mask that removes pad tokens from attention, from the mask draw, and from the loss. We tried giving it more, a 16-d summary of the uncovered frames (their whitened mean plus an indicator bit, 264→64→16) projected into the global token; it buys no fidelity and erodes deletion, so the deployed model omits it.

Training. The three modules train from scratch and chain rather than jointly (Fig. 3): the converter extracts an action-word set from every training clip, then the prior and the decoder train on those same extracted sets, so absence is a trained condition (DG4). Each module trains on a single GPU over the HumanML3D training split. The prior *structure* network (bidirectional transformer, d=384, 6 layers) trains with AdamW (lr 2×10^{-4}, weight decay 0.01, batch 256) for 100k steps under MaskGIT random-subset masking, reconstructing the masked slots by

$$\mathcal{L}_s \;=\; \mathrm{CE}(\pi) \;+\; \mathrm{CE}(g) \;+\; \mathcal{L}_{\mathrm{diff}}(\mu^0, w^0), \tag{5}$$

its window heads small per-slot diffusion models (cosine schedule, x_0 parameterization) carrying the $\mathcal{L}_{\mathrm{diff}}$ term; the prior *content* regressor (same size) trains under the same optimizer and budget by the direction-plus-magnitude loss

$$\mathcal{L}_y \;=\; \big(1 - \cos(\hat{y}_i,\, y_i)\big) \;+\; \big\|\hat{y}_i - y_i\big\|_2^2, \tag{6}$$

teacher-forced on the converter structure; the two prior networks are the same non-causal decoder blocks (slot self-attention plus cross-attention to the caption tokens, the decoder's *memory*), differing only in masking and objective. The decoder (d=512, 8 layers) trains with AdamW (lr 1×10^{-4}, OneCycle schedule, 5% warmup, batch 64) by masked-token denoising with span-jitter and compose augmentation. Each example draws its mask from a fixed schedule: a random token subset at a cosine ratio (45%), the spans of one to three action words (30%), or the full clip (otherwise); independently, 5% of examples zero the set's presence, training the $\varnothing$ of Eq. 7. No geometry stage optimizes its own objective: the refiner is learned end-to-end through the splatted logit (Eq. 4).

Inference. At sampling (§3.4) decoding is one MaskGIT unmasking pass (pass-count study in Table 2), amplified by classifier-free guidance over the *whole* set,

$$\hat{\epsilon} \;=\; \epsilon_\phi(x_t, t, \varnothing) \;+\; w\,\big(\epsilon_\phi(x_t, t, U) - \epsilon_\phi(x_t, t, \varnothing)\big), \tag{7}$$

with $\varnothing$ the presence-zero set that complete-pair training makes in-distribution. When the set comes from the caption pipeline its regressed content is *snapped* to the nearest training vector, a train-free step recovering two-thirds of the pipeline FID gap (Appendix D); residual jerk at patch seams is removed by fusing two half-patch-offset decodes each step (MultiDiffusion (Bar-Tal et al., 2023) on time), bringing jerk to the mocap floor at no cost (Table 2).

C Parameter Studies

The dials that matter. ours exposes the dials 3DGS does; Table 2 reports the three against the same full-test FID. The action word budget N prunes quality gracefully to a floor; guidance w serves generation and editing at once, FID U-shaped while edit obedience stays flat-to-rising; and the decode head, not the action words, owns the jitter, the joint patch denoiser halving foot skate and FID in-model with a foot-contact term driving action words-suffice foot skate to 0.069, below GT. Collapsing the five tracks to one costs FID (+0.014) and foot skate, so the concurrency coordinate does real work, and giving the decoder a summary of the uncovered frames (Appendix B) buys nothing while eroding deletion (0.385 vs. 0.333), the capacity asymmetry DG4 predicts.

Table 2: **Parameter studies** (full-test FID unless noted). Quality is flat in decode passes K (one-shot), so K is not swept here.

Dial	Setting swept	Effect
action word budget N (opacity prune)	$N{=}2/4/8/16$	FID $1.50/0.44/0.13/0.105$ (floor 0.099)
Guidance w	$0 \to 3$	FID U-shaped, edit obedience flat-to-rising
Decode head	per-frame / joint patch	FID $0.058 \to 0.033$, foot skate $0.25 \to 0.101$; foot-contact term a further $\to 0.069$

Where the jerk lives, and the noise-free fusion that cancels it. A diagnosis localizes 93% of the excess jerk to the patch seams, where independent per-patch draws inject uncorrelated noise. Dropping the per-patch tail injection there takes jerk from 1.64 to 1.31; the residual is cross-seam draw mismatch, so we decode two half-patch-offset grids and fuse their states at every denoising step (the MultiDiffusion move on the time axis). Fusion reaches jerk 0.65, the mocap floor (0.66), with FID $0.041 \to 0.040$, foot skate $0.059 \to 0.057$, and R@3 unchanged over three seeds; the two grids decode the same motion (TMR cosine 0.998 between them), so fusion cancels seams rather than averaging modes, at a mild cost of acceleration tails $\sim 11\%$ below mocap. Retraining does not remove the need: an in-model jerk penalty trained from scratch drives the free decode to baseline jerk (1.00 at FID 0.043, R@3 0.797, 750k steps), still short of the fusion floor, so the inference-time fix beats retraining. The editing referees do not see the dial: switch obedience and deletion residual are unchanged across the single, blended, and fused decodes (TMR cosine 0.999).

D THE MANIFOLD SNAP

The caption→action words prior and the motion→action words encoder must agree distributionally to decode well. On one fixed the prior structure (matched seed, one-shot decode) we vary only how content is produced (Table 3). Structure is not the gap: action word-count, track, center, and width marginals match the converter extraction (Jensen-Shannon ≤ 0.003 on count and track). Content is: the regressor's vectors sit off the training manifold (nearest-neighbor cosine distance $0.097 \to 0.120$), mode-average within clusters (spread $0.253 \to 0.221$), and decode at 0.176. Snapping each vector to its nearest training neighbor, train-free, closes two-thirds of the gap to the 0.033 extraction ceiling (0.082); a learned content head lands nearer the manifold (0.127) but snapping it (0.087) still trails regressor-plus-snap. Every conditioning-corruption or post-hoc content repair we audited failed with a measurable mechanism, so the recipe is: amplify the condition (set-null guidance, §3.4) and snap the content; do not corrupt it or edit the samples.

Table 3: **The pipeline FID gap is content, not structure** (prior structure fixed, matched seed, full test): the regressor lands off-manifold (0.176), the train-free snap recovers most of the gap to the 0.033 extraction ceiling (0.082, R@3 0.842), and a learned content head (0.127) does not reach it. Full-sampling pipeline: 0.092 (Table 4).

content mechanism (fixed the prior structure, $K{=}1$)	FID↓
the converter extraction (set-only ceiling)	0.033
the prior regressor (naive)	0.176
the prior generative content head	0.127
generative head + manifold snap	0.087
the prior regressor + manifold snap (deployed)	**0.082**

E GENERATION QUALITY AND EVALUATION PROTOCOLS

Generation quality. Generation is measured on the full HumanML3D test set over three runs through one byte-identical harness, with MDM (Tevet et al., 2023), MoMask (Guo et al., 2024), and SALAD (Hong et al., 2025) re-measured from released checkpoints so every row is like-for-like; FID, R-precision (true caption among 32 candidates), and MM-Dist use the Guo et al. (2022) evaluator. Decoding each clip from its own action words with no text reaches FID 0.040 at R@3 0.798, below every measured baseline (Table 4); this quality read decodes the *full clip* from the set—edit-time re-rendering stays windowed (§3.5)—and the caption pipeline holds R@3 0.843 at FID 0.092, the gap the prior's sampled *content* not its placement, two-thirds recovered by the snap

(Table 3). The set-only re-decode leads every measured FID; the caption pipeline trades FID (0.092 vs. MoMask's 0.057, SALAD's 0.086) for retrieval (R@3 0.843, the highest measured).

Table 4: **Standard-suite quality** (full test; ours = three from-scratch seeds, K=1; baselines re-measured). Foot-skate beats GT and every baseline; jerk at the mocap floor.

	conditioning	FID↓	R@3↑	MM-Dist↓	FS↓	jerk↓	bone↓
GT reference (harness ceiling)	–	–	0.806	2.94	0.073	0.66	**0.00**
MDM (Tevet et al., 2023) (measured)	caption	0.476	0.711	3.66	0.082	0.74	0.48
T2M-GPT (Zhang et al., 2023a) (measured)	caption	0.171	0.755	3.28	0.079	1.29	0.80
MoMask (Guo et al., 2024) (measured)	caption	0.057	0.822	2.95	0.072	1.02	0.63
SALAD (Hong et al., 2025) (measured)	caption	0.086	**0.865**	**2.65**	0.080	0.87	1.05
Ours, action words-suffice (action words only)	action-word set	**0.040**	0.798	2.98	**0.057**	**0.65**	**0.28**
Ours, pipeline (caption→action words→motion)	caption	0.092	0.843	2.74	0.088	0.74	0.30

Referees. Paired-preference within-clip ceiling 0.840 (n=487); FID/R-precision via Guo et al. (2022); editing referee the frozen TMR (Petrovich et al., 2023)—the same towers our content space freezes, which is why every edit reading is relative and the coupling is audited below, not assumed away. Switch realize is the donor-lift hit, the same reading as insert.

Table 5: **Preserve, the geometric companion to Table 1's *sem.* column**: joint error on the frames outside the target span (cm; for move, outside $[t_s, t_e{+}\Delta]$); *sem.* asks whether the rest still reads as its actions, this table asks how far its frames moved. The shared edit window is the *group* span, wider than the judged span, so its re-rendered margin is charged to every windowed method alike; ours re-decodes the whole clip from the edited set, its move carries the followers' intended retime, and its delete (8.7) removes every same-content instance. "–" as in Table 1.

	Insert	Switch	Move	Delete
MotionLab	1.8	8.0	1.4	7.9
MotionReFit	20.9	19.2	–	19.0
given the target window	4.2	6.2	–	6.3
SimMotionEdit	60.7	55.6	–	55.5
MotionCLR	23.2	18.7	20.5	17.8
MDM inpaint	1.6	1.9	1.3	1.5
CondMDI	0.0	0.9	–	0.9
Paste-and-regen	2.3	4.8	2.2	1.5
+ smoothing	2.7	4.9	2.4	1.8
Action words (ours)	3.2	5.2	2.2	8.7
on generated (the prior) sets	1.6	5.0	1.9	8.5

Referee–content-space coupling. Both use the frozen TMR towers. Nothing in training maximizes a TMR score (the decoder has no text or TMR channel, DG4), and one harness applies the same readings to every method. This does not by itself rule out a residual preference—TMR could still read some methods' output style more favorably than others—so the claims that matter are cross-checked off-TMR: seam and skate are model-free, FID/R-precision use the separate Guo et al. (2022) evaluator, paste-and-regen carries no content vector yet scores comparably on realize, and the supplementary video plays every showcase case for direct inspection.

Separating the representation from the decoder. The contribution could hide in either half, so the benchmark carries controls that hold one fixed. (i) *Same information, no representation*: paste-and-regen receives our windows and donor phrases but keeps no action-word set—given *where*, generation-and-pasting alone realizes 0.83–0.87, and everything the set adds shows up in what the paste cannot buy (boundary, physics, neighbor semantics; §4.3). (ii) *Same decoder, sets from text*: the prior row edits caption-generated sets with the identical decoder, within 0.14 of the extracted-set realize—the decoder is not fit to one converter's output. (iii) *Same decoder, degraded representation*: relocating a single word instead of its group drops move 0.76 → 0.30 (Table 7 sweeps the rule)—structure inside the representation, with the decoder untouched, carries the edit.

F Emergent Action Cores

The action word's temporal geometry is learned by the generation objective alone, with no impact supervision, yet its center lands on the moment a human calls the action's core. Sorting action words by *sharpness* (peak-to-mean track energy over the span) shows the learned Gaussian narrowing monotonically: median σ (as a fraction of a quarter-span) falls from 1.89 in the softest quartile to

1.23 in the sharpest. Table 6 recomputes the learned center's agreement with held-out human impact annotations (Jung & Lee, 2026) at the *stroke* level: on the HumanML3D test split we match each annotated stroke to the same-track action word by span IoU (≥ 0.3) and compare μ to the impact frame. The center is contained in the human boundary at least 97.6% of the time per class (100% on five of seven); against the velocity-peak baseline, μ improves on four of seven classes, with kick and jump clean wins, punch ill-defined and only tying, and throw and hop trailing, so we claim only kick and jump (§E). The referee stays held-out: of $2{,}502$ non-mirror strokes only 326 fall on the test split with a valid track and impact frame (278 surviving the IoU gate); the other $1{,}884$ are train-split.

Table 6: **Emergent-core boundary agreement** (held-out HumanML3D test, learned μ vs. human impact annotations (Jung & Lee, 2026)). The learned center is contained in the human boundary at least 97.6% per class and beats the velocity peak on kick and jump; no impact label or velocity peak trains a module. Rendered per class in Fig. 11.

Class	n	$\lvert\mu - \text{impact}\rvert$ med.	vel.-peak med.	$\mu \in$ boundary
kick	90	1.2 f	4.0 f	100%
jump	72	0.9 f	2.0 f	100%
punch	45	4.4 f	4.0 f	97.8%
throw	42	1.7 f	1.0 f	97.6%
catch	15	1.8 f	2.0 f	100%
clap	13	4.0 f	5.0 f	100%
hop	1	5.7 f	0.0 f	100%

G The Action Group

Whole-body edits act on the *action group*, the target action word plus every span-overlapping action word, with no learned grouping. The rule is purely geometric; Table 7 sweeps its definition on the n=128 delete protocol (deployed decoder, w=2, converter false-fire base 0.10) and trades coverage for precision monotonically. Any narrowing leaves partners that re-assert the action, so the widest rule (span overlap, IoU > 0) is the default: it is the only setting under which relocation and deletion move the whole action rather than a fragment of it. Two co-occurring *different* actions do not fuse: on top of the span rule, switch and delete act only on the target's *same-content* members, so waving while walking, the wave switches or vanishes and the walk words stay (the boxing cleanup of Fig. 12 deletes the kicks under surviving punches exactly this way); *move* deliberately does not separate them—the co-spanning slots shift as one beat, since a wave moved off its steps is a different scene, and single-word relocation is the 0.30 ablation above.

Table 7: **Action-group rule ablation** (n=128 delete protocol, deployed decoder). The column is delete *residual* (leftover adherence to the removed action, lower is better), not the delete *realize* of Table 1; the deployed whole-group rule (residual 0.375) corresponds to that table's realize 0.94. Narrowing the group trades coverage for precision; any narrowing leaves partners that re-assert the action, so the widest overlap rule is the default.

Group rule	avg. group size	delete residual
span overlap (IoU > 0, default)	4.9	0.375
IoU ≥ 0.3	3.4	0.625
IoU ≥ 0.5	2.5	0.758
same-span (center-matched)	2.7	0.758
single action word	1.0	0.883

H Qualitative Panels

The panels below back claims made in the body: the substrate comparison (§3.1), the action-group vs. single-word retiming (§H.1), and the emergent core per class (§E). All SMPL decodes are from the clip's own action words with no caption unless noted.

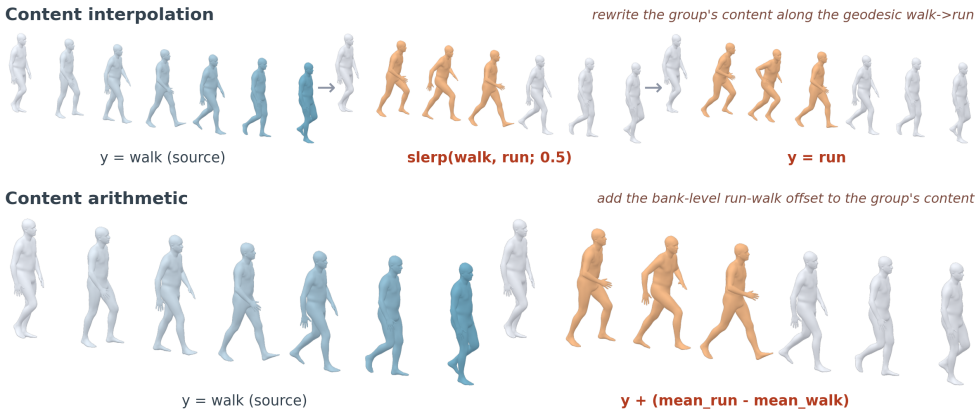


Figure 8: **The content coordinate is open** (one captured walking clip; the edited group's span renders warm, untouched frames stay pale). *Top—interpolation*: the group's content slerped toward a bank *run* unit at $\lambda = 0, 0.5, 1$; the arm carriage lifts and the stride quickens through a plausible midpoint (span root speed $0.6/3.9/10.6$ cm/f; frozen-TMR read-back turns to *run* from λ=0.5). *Bottom—arithmetic*: adding $\bar{y}_{\text{run}} - \bar{y}_{\text{walk}}$ (bank means; no donor unit) to the same group flips the read-back ($-0.46 \rightarrow +0.36$): vector arithmetic on the action coordinate.

H.1 Composition and attribute edits

Beyond the four benchmark operations, Fig. 6 shows *compose*, *track swap*, *manner*, and *resize*. *Compose* unions action words across tracks (a both-arm raise, a mirror pair on the two arm tracks, over a squat): on a *different* body-part track ours keeps both actions alive 80% with the donor retrievable, where STMC (Petrovich et al., 2024), regenerating rather than editing, keeps both alive only 36%. A mirror-mapped *track swap* moves an action to the other leg (70% of one-leg actions, a bare swap flips none), a *manner* residual raises the jump with identity kept (Appendix E), and *resize* (the jump group's windows $\times 1.7$, later action words retimed by the extension; same clip as *manner*) lengthens the airtime ($8 \rightarrow 12$ airborne frames) and carries the jump almost a meter further (0.9 m); non-ballistic actions barely change, so a clean resize is future work.

H.2 The open content coordinate: interpolation and arithmetic

Design goal DG2 keeps the content coordinate continuous, and the coordinate cashes that promise directly (Fig. 8); every number below is measured on the decoded clips. *Interpolation*: on a captured walking clip, rewriting the group's content along the geodesic toward a bank *run* unit morphs the gait through a plausible middle—root speed in the span rises monotonically ($0.6 \rightarrow 3.9 \rightarrow 10.6$ cm/frame at $\lambda = 0, 0.5, 1$) and the frozen judge's read-back flips from *walk* (cos margin -0.46) to *run* already at λ=0.5 ($+0.38$; $+0.64$ at λ=1). *Arithmetic*: adding the bank-level offset $\bar{y}_{\text{run}} - \bar{y}_{\text{walk}}$ to the same group—no donor unit involved—flips the read-back the same way ($-0.46 \rightarrow +0.36$) with a milder gait change, the analogy-offset behavior that word embeddings made familiar (Mikolov et al., 2013). A codebook exposes neither write: both are points a discrete inventory does not contain; frames outside the group's span keep walking untouched (DG1).

I Application Scenarios on Captured Clips

Fig. 1 shows the authoring side: a caption becomes a set and the set is revised. The three sessions below run the same writes on *captured* clips (held-out HumanML3D), with no caption anywhere: the converter reads the clip's action-word set off the raw motion (§3.2), the write edits the set, and the clip re-decodes from the edited set: the untouched action words re-render the capture's own actions in place (§3.5). Timeline lobes are the learned splats, colored by content; the legend under each figure names each color's retrieval read-back, and grey lobes are incidental read-backs the edit does not touch.

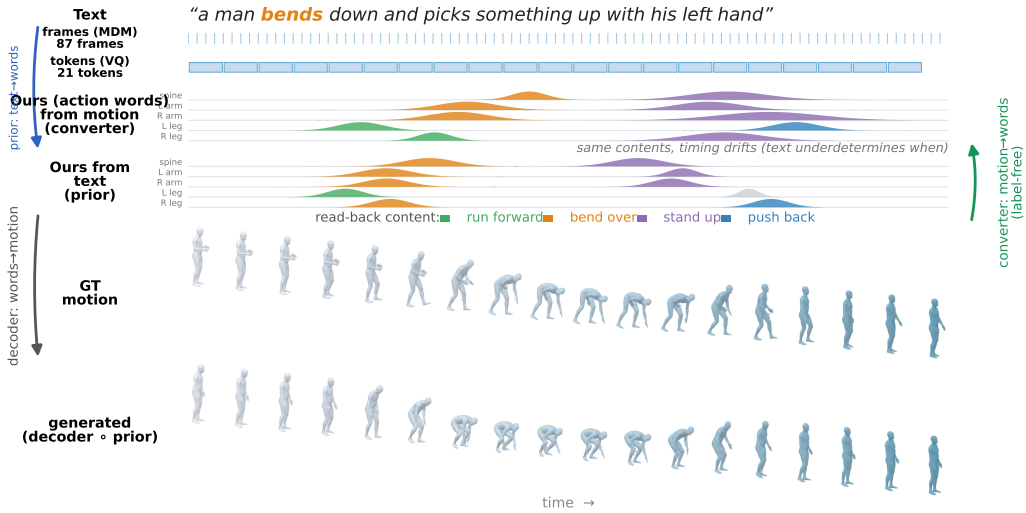


Figure 9: **One clip, three substrates, and the round trip** (held-out clip M003592). Frame (MDM) and token (VQ) models place the two actions on all 87 frames or 21 uniform tokens; ours reads the same clip as action words on body-part tracks (the converter), the gap between actions carrying none. Below, the caption samples a set (the prior) with the same contents but drifting timing, and SMPL strobes show the GT clip and the prior decode. Labels are retrieval read-back.

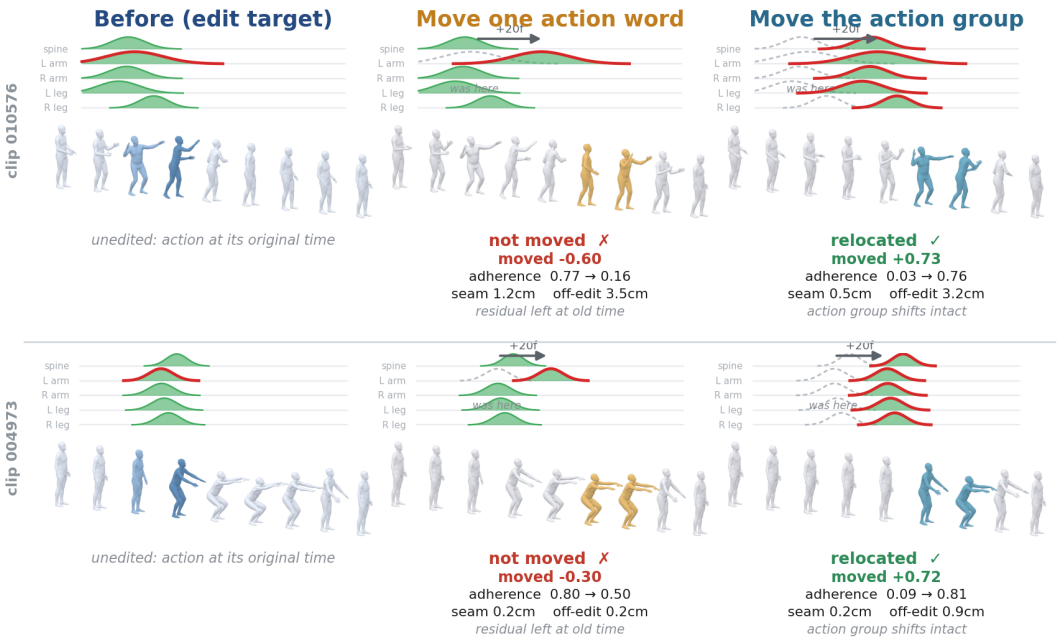


Figure 10: **Retiming: single action word vs. action group** (a throw and a squat clip). Above each strobe a five-lane action word timeline (spine, arms, legs) tracks the action group: *Move one action word* shifts only the target action word (red) and leaves its co-spanning neighbors at the old time, so the action tears and barely relocates (moved <0, adherence collapses); *Move the action group* carries every co-spanning action word to the new cue together (moved $\sim+0.7$). We act on the action group by a fixed rule; learning a finer edit action word is left to future work (§5).

Cleanup: drop the kicks (Fig. 12). A captured boxing-form take mixes kicks into the punching the director wants kept. Because the converter reads the take back as action words, the kicks are addressable: one *delete* removes every same-content instance with its group, and the decoder bridges the vacated window from the remaining set—absence is a trained condition—while every frame outside the window stays the original capture.

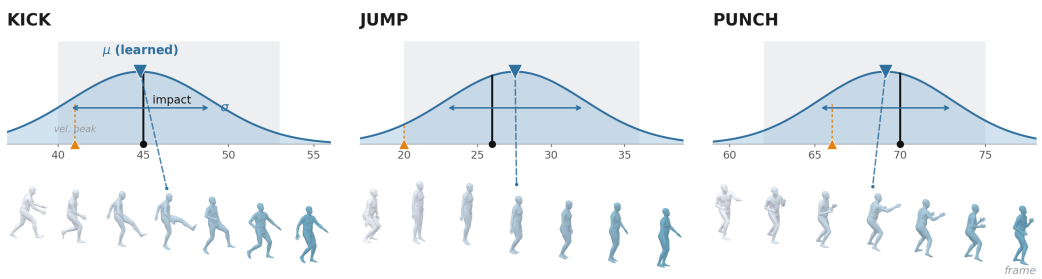


Figure 11: **Emergent core, one test clip per class** (blue ▼ = learned μ, black = the human impact annotation, orange ▲ = the velocity-peak *kinematic baseline*, not an annotation, arrow = σ). Kick and jump win at the class level; punch is kinematically ill-defined and only ties the peak. No impact label or velocity peak trains a module (Table 6).



Figure 12: **Cleanup on a captured clip.** The converter reads the take's action-word set (timeline above each strip); one *delete* write removes the kicks—every same-content instance, its whole group (dashed ghosts)—and the decoder bridges the vacated window from the remaining set. Every frame outside the window is the original capture.

Timing pass: pull the middle beat earlier (Fig. 13). In a run–vault–return capture the mid-take crossing lags its cue. One *move* pass pulls the beat's action words -12 frames—the crossing's group and the overlapping run-back words around it travel together, so the jump itself is untouched—the landing re-integrates at the new time, and the approach run before the window is the original frames.

Variant iteration: one more, then re-strike (Fig. 14). Blocking a karate combo, the choreographer first wants *more*: one write repeats the kick group's own words a beat later, and the double kick becomes a triple. On review, the opener changes character instead: *switch* lifts the first kick's group out and writes a punch group—the punch's own words, on its own tracks—at the same beat. Each variant is one write to the same clip—the rest of the take persists across the iteration—so options accumulate on one timeline rather than through re-capture or re-prompting.

J Limitations, Expanded

The boundaries stated in §5, with the evidence behind each and what already works inside it.

Group granularity. The action group is a fixed overlap rule, and it is the right default: relocating a single action word instead of its group drops move from 0.76 to 0.30 (Fig. 10; Table 7 sweeps the same rule on delete)—co-spanning parts of one action must travel together. The cost is per-part freedom *within* a whole-body action (retime only the arms of a jump); a learned, finer edit unit that factorizes the tuple into a content core plus per-track residuals (§5) is the natural next step, and

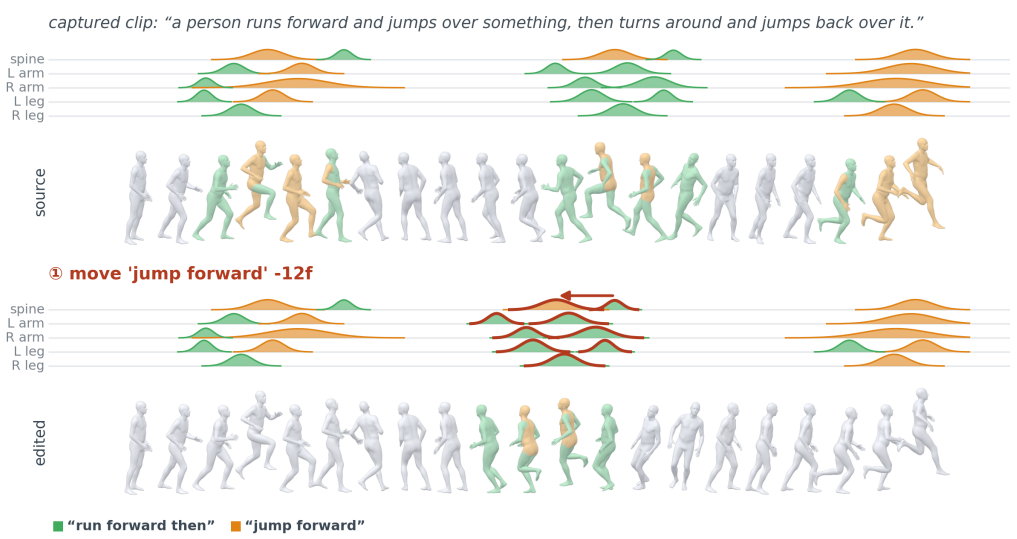


Figure 13: **Timing pass on a captured clip.** The mid-take crossing is pulled 12 frames earlier by one *move* pass (arrow); the crossing's group and the overlapping words around it travel together, and the landing re-integrates at the new time.

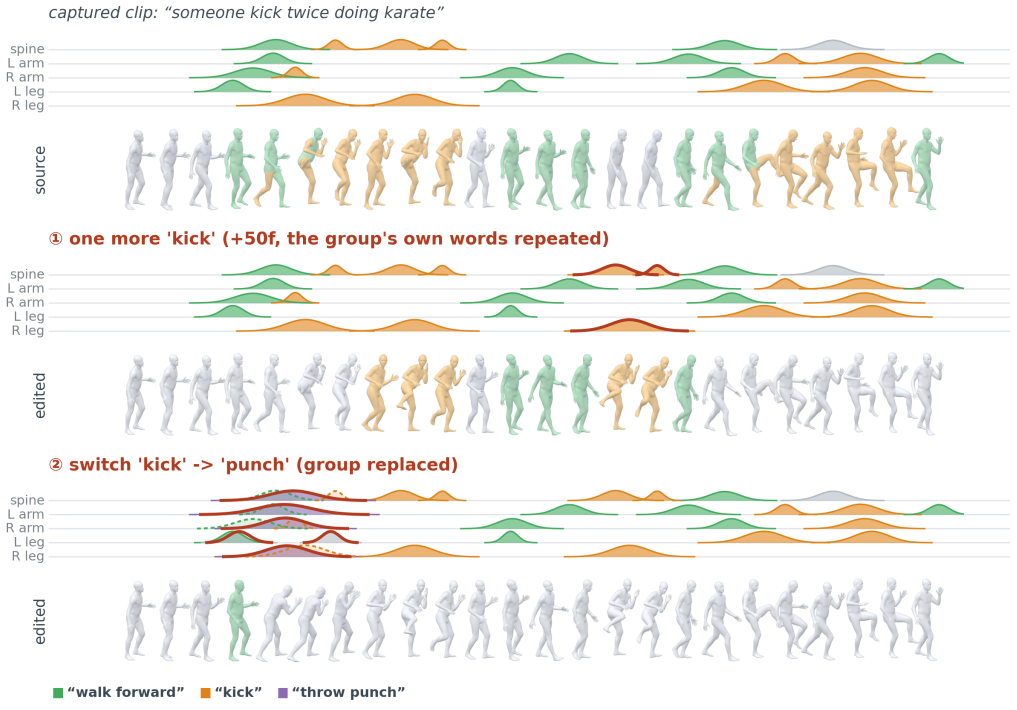


Figure 14: **Variant iteration on a captured clip.** Two takes on the same combo: *one more* (the kick group's own words repeated a beat later—the double kick becomes a triple) and *re-strike* (the first kick's group lifted out, a punch group written in its place, the punch's own words on its own tracks). The rest of the take persists across both.

track-level writes on actions that already live on one track (track swap, per-track compose; Fig. 6) work today.

Attribute edits are graded, not uniform. *Manner* transfers reliably along verified axes (speed d=1.10, jump height d=1.15, stance d=−1.01); *resize* cleanly extends ballistic actions (the jump gains 4 airborne frames and 0.9 m) while sustained actions barely lengthen—their duration lives

in content, not in the window alone—and the extension is kinematic, not dynamic: the decoder stretches the airborne phase without re-solving ballistics, so a strongly resized jump reads as reduced gravity, which the contact metrics do not see; and *compose* keeps both actions alive 80% across tracks (vs. STMC's 36%) but is the noisiest write on same-track overlaps. None of these blocks the four benchmark operations; they mark where the attribute coordinates are coarser than the membership ones.

Traversal control. The set fixes *what* happens *when*, on *which* body parts; where the body travels while doing it—distance covered, heading held or turned—is inferred by the decoder from content and context, not exposed as a coordinate. In long multi-action re-renders the inferred path can drift or turn between actions even when each action itself is correct, and no set write can pin it. A traversal coordinate on the word (or a root-path channel beside the set) is a natural extension, and it would also give the benchmark a direct path-error axis, which the frozen retrieval judge cannot see.

Clip horizon. Everything here is trained and evaluated at HumanML3D's ten-second cap with $M{=}24$ slots (corpus P99). The set representation itself is length-agnostic—the slot count and the clock are the only dials—but the minutes-long scene needs hierarchical grouping (scenes of groups of action words), a streaming decoder, and interaction design at the timeline, none of which this paper evidences.

K Structural Motion-Editing Benchmark: Example Cases

Each case is a held-out HumanML3D clip, a named target action in it, and one action-group operation. Switch, move, and delete each run on all 140 clips; insert runs on its own 88-clip set (from 120 candidate gaps after the motion-idleness audit). Cases are balanced across the seven action classes and released with the code.

The smoothed row. Per boundary and per op we sweep inertialization (quintic C^2), a boundary cross-fade, and a Gaussian temporal filter ($K \in \{3, \dots, 12\}$) in feature space, and report the seam-minimizing configuration (cross-fade $K{=}10$ for insert, Gaussian $K{=}12$ otherwise). The boundary jerk collapses (7.2–41.2 $\rightarrow$ 0.1–1.8 cm); the trade remains: insert realize falls $0.86 \rightarrow 0.81$ (the blend blurs the entry), and foot-skate rises above the un-smoothed rows (insert $.084 \rightarrow .111$, switch $.081 \rightarrow .124$)—the seam-minimizing configurations are also the skate-worst ones.

Reading the seam against the GT band. A lower-is-better jerk metric also rewards over-smoothing, so the seam is read as a ratio—peak boundary jerk over the clip's interior median $|\Delta v|$—against the band the same statistic occupies at real action entries: $[1.8, 12.5]\times$ (median 3.5). Ours sits inside the band on the landing operations ($3.3/3.8/2.6\times$); the smoothed paste sits inside it on switch and move ($4.7/1.9\times$) and falls below only on insert ($0.4\times$); the rewrite editors sit at the band floor (1.2–$2.2\times$); the raw paste spikes far above it. Delete is a removal, so its low bridge ratio ($1.5\times$) is the correct reading there.

Table 8: **Seam as a ratio read against the GT action-entry band.** Peak boundary jerk over the interior median $|\Delta v|$, median over cases. GT band (real action entries at the edited span) $= [1.8, 12.5]\times$ (median 3.5). On the landing operations ours sits inside the band; the smoothed paste falls below the floor only on insert ($0.4\times$) and sits inside otherwise; the rewrite editors sit at the floor; the raw paste spikes above it (tear). Delete is a removal, so a low bridge ratio is correct there. "–" = operation not expressible / not scored (as in Table 1).

	Insert	Switch	Move	Delete
GT action entry (band) (band $[1.8$–$12.5]\times$)	3.5	3.5	3.5	–
Paste-and-regen	165	54	11.0	14.7
+ smoothing	0.4	4.7	1.9	2.6
Inpaint (MDM)	12	16	6.7	16
MotionCLR	1.2	1.7	1.6	1.8
MotionLab	–	1.4	1.6	1.5
MotionReFit	–	2.0	–	1.9
SimMotionEdit	–	2.2	–	2.1
Action words (ours)	**3.3**	**3.8**	**2.6**	1.5

Ours: edit mechanics. The commands in Table 11 are descriptive labels, not model input: ours receives a modified action-word ledger (span, track, donor content, timing delta), so no method

Table 9: **Plausibility, second axis: bone-length jitter** (mean over bones of the temporal std of bone length, cm), per op, on the identical outputs as Table 1, whose *skate* columns already carry the foot-skate axis; rows follow its family grouping. A rigid skeleton scores 0; a feature-space blend stretches limbs—smoothing a paste-and-regen produces the largest insert/switch bone jitter of any automatic editor (0.71/1.06), while its move and delete fills carry the sampler's own jitter (0.6–0.7). †outputs re-fitted to a rigid template skeleton, ≈ 0 by construction.

	bone-length jitter (cm)↓			
Method	Insert	Switch	Move	Delete
Instruction-tuned				
MotionLab	0.38	0.37	0.40	0.37
MotionReFit†	0.02	0.00	—	0.00
SimMotionEdit†	0.00	0.00	—	0.00
Attention manipulation				
MotionCLR	0.70	0.64	1.19	0.56
Masked infilling				
MDM inpaint	0.64	0.77	0.69	0.71
CondMDI	0.20	0.42	—	0.41
Paste-and-regen				
Paste-and-regen	0.53	0.77	0.61	0.71
+ smoothing	**0.71**	**1.06**	0.72	0.74
Action-word set				
Action words (ours)	0.32	0.37	0.28	0.19
on generated (the prior) sets	0.25	0.38	0.24	0.21
Source re-decode (no-op; the floor)	bone 0.27 cm (FS .033; n=140 source clips)			
Raw capture, same detectors	bone 0.00 cm (FS .055)			

reads the instruction sentence. Every target span is *extracted* by the converter, not invented by the harness. Each op mutates the action group $\mathrm{slots}(t_s, t_e)$ under the content gate of §3.5: *insert* adds one slot into an interval empty on all tracks (≥ 24 frames); *switch* replaces the target's same-content members with the donor group on its own tracks (co-spanning unrelated words—locomotion under an arm action—stay); *move* shifts the co-spanning slots as one beat (Δ=20; single-word relocation is the 0.30 ablation, Fig. 10); *delete* zeroes the target's same-content members plus every same-action instance on the clock (cosine > 0.60, or > 0.50 with the same head verb; span-only deletion leaves refill sources—whole-group lifts realize $0.77 \rightarrow 0.94$ extracted, $0.66 \rightarrow 0.81$ generated). Delete's wider scope is not a judging asymmetry: the referee crop is the same $[t_s, t_e]$ for every method; the scope is priced in preserve (8.7 vs. 0.9–1.8 cm, Table 5), and the window-scoped variant of ours is exactly the span-only ablation.

The shared edit window. Every window-consuming baseline (paste-and-regen, MDM inpaint, CondMDI, and MotionReFit's given-window variant) operates on one shared window: the target *group's* span union (every present action word overlapping the target span)—the same element ours edits, read directly off the representation. Judging crops remain the target span $[t_s, t_e]$ for every method, ours included.

How each baseline realizes an edit. All baselines inherit the identical clips, windows, and donor phrases.

- *Paste-and-regen* copies every frame outside the window verbatim and fills the window per op; the source caption is not used. *Insert* and *switch* sample MDM from the donor phrase alone at window length ($\geq$32 frames, CFG 2.5, time-resampled to the span; Table 10) and paste the sample full-body—a variant pasting the donor's GT window instead scores the same realize with the same tear, so the deployed arm generates. *Delete* inpaints the window away text-free (empty caption, guidance 1.0); *move* copies the window's frames Δ later and fills the vacated remainder the same way. Delete alternatives, for the record: a linear bridge leaves not-yet-returned poses as leftover motion, and refilling from a whole-clip regeneration keeps re-drawing the named action (0.71) unless the caption is surgically stripped (0.94, at a 44 cm tear, failing on 40% of captions)—the text-free fill needs neither.
- *Inpaint (MDM)* locks every frame outside the shared window in the sampler's mask and re-denoises only the window from text (guidance 2.5; insert uses the donor phrase alone as its caption, its best-performing conditioning, $0.34 \rightarrow 0.57$ vs. the full caption). Its *delete* runs

the same text-free infill as paste-and-regen's, and its *move* regenerates the *shifted* window without vacating the original, so the source action persists.

- *CondMDI*, a trained text-plus-keyframe in-betweener, observes the locked surround and infills the shared window under the target/donor caption—the editing construct proper—for the single-window ops: insert, switch, and delete (the delete infill runs *text-free*, as paste-and-regen's); move needs two windows and is left blank rather than faked.
- *MotionCLR* takes no explicit boundary: it steers attention inside its own fresh caption-generation, not a source-conditioned $A \to B$, hence its preservation error; its (de-)emphasis weights were probed per op on 16 cases and deployed at each op's best.
- *MotionLab* runs the released `motionflow` checkpoint. Its outputs move but are instruction-insensitive (changing the instruction moves donor similarity by ≈ 0.001; its delete 0.60 is reproduced by grading its switch outputs as deletions)—consistent with the authors' advisory that the checkpoint predates training-data fixes.
- *SimMotionEdit* (Li et al., 2025b) runs its released checkpoint through the same SMPL bridge; like its family it regenerates the whole clip, near-stationary and source-unrelated (delete trivially 0.94, preserve 55–61 cm). OmniME, TM-Edit, and MotionAgent released no runnable code and are cited only.
- *MotionReFit* runs its released auto-regressive model (canonicalized 16-frame 10 fps chunks); *move* is inexpressible (no span-scoped mode, no in-betweening). The *given the target window* row restricts its whole-clip output to the shared window with the source outside—a generous variant, since the model itself accepts no window.

The sampler's short-window regime. Sampling MDM directly at the target length collapses below the dataset's 2-second minimum (phrase hit $0.55 \to 0.19$ by $L{=}16$; Table 10); sampling at 40 frames and time-resampling to L (the deployed rule, MINL=32) restores the phrase at a foot-skate cost that grows with compression. The shared windows have median 28 frames (86% below 40), so the paste's insert/switch ceiling (§4.3) is a property of short-window generation, and the resample rule is already its best operating point.

Table 10: **Short-window fidelity of the base sampler** (MDM, the paste's generator): 140 donor phrases sampled at length L, scored against their own phrase by the frozen TMR. *Exact* samples directly at L; *resampled* samples at 40 frames (the dataset minimum) and time-resamples to L—identical above 40. Exact-length sampling collapses below 2 s; resampling recovers the phrase at a foot-skate cost that grows with compression. The benchmark's shared edit windows have median 28 frames.

	Exact length			Resampled from 40		
L (frames)	cos	hit@.5	skate	cos	hit@.5	skate
16	0.26	0.19	.069	0.47	0.54	.364
24	0.36	0.36	.067	0.47	0.55	.179
32	0.44	0.49	.065	0.47	0.54	.080
40	0.47	0.55	.061	= exact		
56	0.48	0.52	.063	= exact		
80	0.48	0.50	.058	= exact		
112	0.47	0.51	.053	= exact		
160	0.48	0.49	.054	= exact		

A representative sample of the released case list:

Table 11: **Benchmark example cases** (§4.2; five of each operation shown); all clips are held-out HumanML3D *test* clips. *Span* is the target window $[t_s, t_e]$ (for insert, the audited empty gap), extracted by the converter and handed identically to every method; the referee scores realize on that span's crop. The *Edit* column is a descriptive label only, not any method's generation caption; ours receives the modified ledger, not the sentence.

Operation	Clip	Source caption (HumanML3D, held-out test)	Span / T	Edit
Insert	008382	the toon has both arms raised at an angle above their head, as to be in the squatting motion for exercise	[96, 122] / 157	insert a throw into the gap
Insert	006652	the man is doing starjumps	[34, 101] / 196	insert a kick into the gap
Insert	010940	a person throws something with their right hand	[57, 84] / 196	insert a punch into the gap
Insert	002139	it looks like he is lifting a box on a shelf and grabbing another box to put on shelf as well	[45, 85] / 167	insert a clap into the gap
Insert	011673	subject is sitting flat on the ground feat straight in front then the subject stands straight up then sits back down with feet straight out in front again	[39, 72] / 196	insert a raise into the gap
Switch	003424	man jumps twice in place	[10, 30] / 94	switch the jump to a kick
Switch	012657	kick left leg step back	[10, 35] / 95	switch the kick to a squat
Switch	010940	a person throws something with their right hand	[112, 127] / 196	switch the throw to a kick
Switch	007767	a person slowly kicks with their right leg	[23, 49] / 96	switch the kick to a jump
Switch	002606	a man is shadowboxing while standing still	[6, 18] / 51	switch the throw to a kick
Move	001592	a person does a small jump	[6, 28] / 87	move the jump +20 frames later
Move	004873	a person throws something with their right hand	[16, 37] / 83	move the throw +20 frames later
Move	011055	a person jumps up and down once	[16, 36] / 65	move the jump +20 frames later
Move	000348	a person kicks their right leg up twice while hopping between feet, then punches their right arm	[115, 133] / 196	move the kick +20 frames later
Move	008801	a man jumps forward with his arms at his sides	[26, 41] / 56	move the jump +20 frames later
Delete	001214	a man hops on his right leg five times, makes a small jump, then steps forward into a strong push with both hands. he then returns to the starting position	[10, 38] / 127	delete the hop
Delete	012459	a person turns steering wheel with both hands	[108, 122] / 196	delete the squat
Delete	004499	a man is doing jumping jacks	[15, 28] / 83	delete the jump
Delete	001781	a man squats deeply three times while raising both arms in the air as if holding a dumbell	[2, 13] / 76	delete the squat
Delete	006686	a man makes two spartan kicks, followed by two medium high kicks	[39, 62] / 196	delete the kick