FROM BEHAVIOR TO KINEMATICS: A HIERARCHICAL MOTION-LANGUAGE BENCHMARK FOR RODENT MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mice and rats are central to experimental studies of brain function and behavior, where brief changes in posture and local movement carry information that broad activity labels can miss. Studying language-conditioned generation of rodent motion requires motion–language pairs that preserve these observed details. We introduce RodentBEK, a benchmark built from real mouse and rat recordings with a shared 25-landmark skeleton and 8,195 motion windows. Its Behavior–Event–Kinematics (BEK) framework connects whole-window behavior captions, half-second event descriptions, and pose-derived kinematic evidence. Events are described from video and summarized into behavior captions; kinematic measurements support verification and revision of movement claims. The benchmark contains 24,585 behavior captions and 46,679 nonempty event descriptions, with source-time-separated training, validation, and test partitions. We evaluate diffusion, autoregressive, and masked-token generators with species-specific motion–text evaluators, comparing behavior-only and behavior-plus-event inputs to the same trained model. Event conditioning improves local event correspondence across all three models and both species, while its effects on global behavior alignment and motion distributions depend on the model. RodentBEK provides a benchmark for studying how generative models use fine-grained language to represent rodent movement.

1 INTRODUCTION

Mice and rats are widely used to investigate how neural activity, genetic variation, and experimental interventions relate to behavior. Posture, locomotion, exploration, and grooming reveal responses to the environment (Gomez-Marin et al., 2014; Datta et al., 2019; Pereira et al., 2020). Computational methods resolve this behavior into short motifs and body-part dynamics from video (Wiltschko et al., 2020; Weinreb et al., 2024; Hsu & Yttri, 2021). This fine-grained view of behavior motivates a generative question: can a model translate descriptions of laboratory rodent behavior into the corresponding motion?

Answering this question requires motion–language pairs that preserve local movement. Brief head turns, forepaw adjustments, and changes in trunk inclination can occur within an otherwise continuous activity. Keypoint-MoSeq resolves subsecond changes in pose dynamics, while B-SOiD identifies behaviors from distances, angles, and speeds of tracked body parts (Weinreb et al., 2024; Hsu & Yttri, 2021). Recording these details requires spatially reliable landmarks and consistent trajectories through occlusion and changing body orientation (Dunn et al., 2021; Li et al., 2023). Language must specify how these movements contribute to the activity. In the rearing example in Figure 1, a whole-clip caption identifies the behavior, whereas local descriptions specify how the head, forelimbs, and trunk move. Pose-derived measurements can then check the physical claims made by those descriptions.

Paired language and motion have supported a broad range of human motion generators (Plappert et al., 2016; Guo et al., 2022). Animal motion generation has expanded species coverage and skeletal diversity through animation-derived resources, including AnimalML3D, Zoo-300K, T2M4LVO, and AniMo4D (Yang et al., 2024; Zhang et al., 2024; Lee et al., 2025; Wang et al., 2025). Kirin's

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.

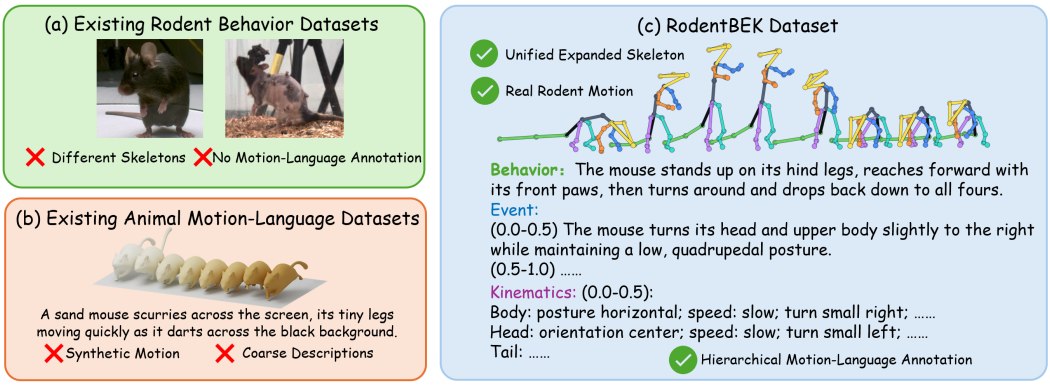

Figure 1: **From existing resources to RodentBEK.** (a) Laboratory pose records and (b) captioned animations provide complementary resources; (c) RodentBEK pairs real recordings with behavior and event descriptions grounded in kinematic evidence.

AiM3D extends the data sources to motion reconstructed from in-the-wild videos (Zhao et al., 2026). Laboratory resources such as Rat-7M and the freely moving mouse recordings of Li et al. (2023) capture real behavior, but provide source-specific pose representations without the paired descriptions required for language-conditioned generation. What remains missing is language paired with laboratory rodent motion at both whole-behavior and local-event scales. Table 1 distinguishes real-animal recordings (Captured), sub-clip language (Localized), and explicit pose-derived evidence (Kinematics). In Table 1, † marks augmented clips, ‡ released-video duration, and Words/text the mean over listed entries. Zoo-300K includes 2,220 clips without released text; T2M4LVO counts seven subject variants per base caption.

Table 1: **Comparisons with public datasets.** RodentBEK combines real laboratory recordings with localized event descriptions, behavior captions, and pose-derived kinematic evidence.

Dataset	Rodent data	Texts	Words/text	Captured	Localized	Kinematics
Zoo-300K (2024)	8,740 clips†	6,520	77.84	×	×	×
T2M4LVO (2025)	25 clips	525	34.56	×	×	×
AniMo4D (2025)	1,285 clips	3,035	10.51	×	×	×
Rat-7M (2021)	2.71 h‡	0	—	✓	×	×
3D Mouse Pose (2023)	3.10 h	0	—	✓	×	×
RodentBEK	8,195 clips	24,585 B 46,679 E	15.23 B 13.98 E	✓	✓	✓

We introduce **RodentBEK**, constructed from the videos in Rat-7M and the mouse recordings of Li et al. (2023). Rodent25 represents both species with 25 anatomical landmarks spanning the head, trunk, limbs, and tail. Behavior–Event–Kinematics (BEK) is a *hierarchical motion–language grounding framework*: B summarizes a complete window, E describes half-second intervals, and K records measured body-part states and changes. Motivated by fine-grained human motion annotation (Wu et al., 2025), the pipeline generates E from video, uses event semantics to assemble windows, and derives three B alternatives from each window's events. K supplies geometric evidence for verifying and revising these descriptions (Figure 2). The resulting 8,195 windows contain 24,585 B captions and 46,679 nonempty E descriptions.

RodentBEK enables us to examine how event-level language affects generated motion. We evaluate diffusion, autoregressive, and masked-token generators trained on both species, comparing B and B+E inputs to the same checkpoint. We measure global behavior alignment, local event correspondence, distribution similarity, and generated variation. E improves local event correspondence

across all three models, while global retrieval and distribution measures show different responses. Our contributions are:

- **RodentBEK**: real mouse and rat motion–language pairs in a common Rodent25 representation, with traceable source-time partitions.
- **BEK construction**: behavior summaries and local event descriptions linked to pose-derived kinematic evidence for annotation verification and revision.
- **Reference evaluations**: three generation paradigms, within-checkpoint B/B+E comparisons, and shared repeated-evaluation protocols.

2 Related Work

RodentBEK connects motion–language datasets and generative models with resources for measuring rodent behavior.

2.1 Motion-Language Datasets

HumanML3D establishes a widely used setting for text-conditioned human motion generation, following earlier paired resources such as KIT-ML (Guo et al., 2022; Plappert et al., 2016). AMASS unifies heterogeneous human mocap through a common body representation, and BABEL adds sequence-level and temporally localized action labels (Mahmood et al., 2019; Punnakkal et al., 2021). FineMotion pairs short snippets with body-part descriptions generated from pose pairs and describes complete sequences with paragraphs (Wu et al., 2025). FineMoGen combines spatial and temporal descriptions for controllable generation and editing (Zhang et al., 2023b).

Animal datasets extend this pairing across species and skeletal structures. OmniMotionGPT introduces AnimalML3D by extracting skeletal motion from DeformingThings4D animations and pairing it with human descriptions (Yang et al., 2024). T2M4LVO uses artist-created motions and alternative short, medium, and long clip captions, with generation conditioned on heterogeneous skeletal configurations (Lee et al., 2025). Kirin introduces AiM3D, which pairs motion reconstructed from in-the-wild quadruped videos with multiple clip captions (Zhao et al., 2026). RodentBEK organizes laboratory rat and mouse recordings into a common anatomical representation, pairing whole-window captions with localized events and structured kinematic evidence.

2.2 Rodent Pose Estimation and Behavior Analysis

DeepLabCut and SLEAP recover user-defined body landmarks from video (Mathis et al., 2018; Pereira et al., 2022). Anipose combines multi-view tracking with geometric and temporal constraints, while anatomically constrained rodent models explicitly represent skeletal kinematics (Karashchuk et al., 2021; Monsees et al., 2022). Rat-7M supplies synchronized multi-view video and motion-capture measurements for three-dimensional kinematic profiling with DANNCE (Dunn et al., 2021). Li et al. (2023) provide multi-view videos of freely moving mice with sparse pose labels and study temporal semi-supervision for markerless tracking, including temporal stability and skeletal consistency. DANNCE and s-DANNCE's graph-based refinement provide the reconstruction basis for RodentBEK (Dunn et al., 2021; Klibaite et al., 2025). We preserve source synchronization and add a common skeleton, motion windows, and language for generation.

Behavior-parsing methods further establish the value of short, local observations. MoSeq identifies recurring behavioral syllables; Keypoint-MoSeq links such structure to tracked pose dynamics; and B-SOiD characterizes behavior using body-part geometry and movement (Wiltschko et al., 2020; Weinreb et al., 2024; Hsu & Yttri, 2021). These studies motivate BEK's local descriptions, which pair source intervals with pose evidence for motion generation.

2.3 Text-to-Motion Generation

MDM generates continuous motion through diffusion (Tevet et al., 2023); T2M-GPT autoregressively predicts quantized motion tokens (Zhang et al., 2023a). MoMask combines residual quantization with masked token prediction and refinement (Guo et al., 2024), while MotionGPT treats

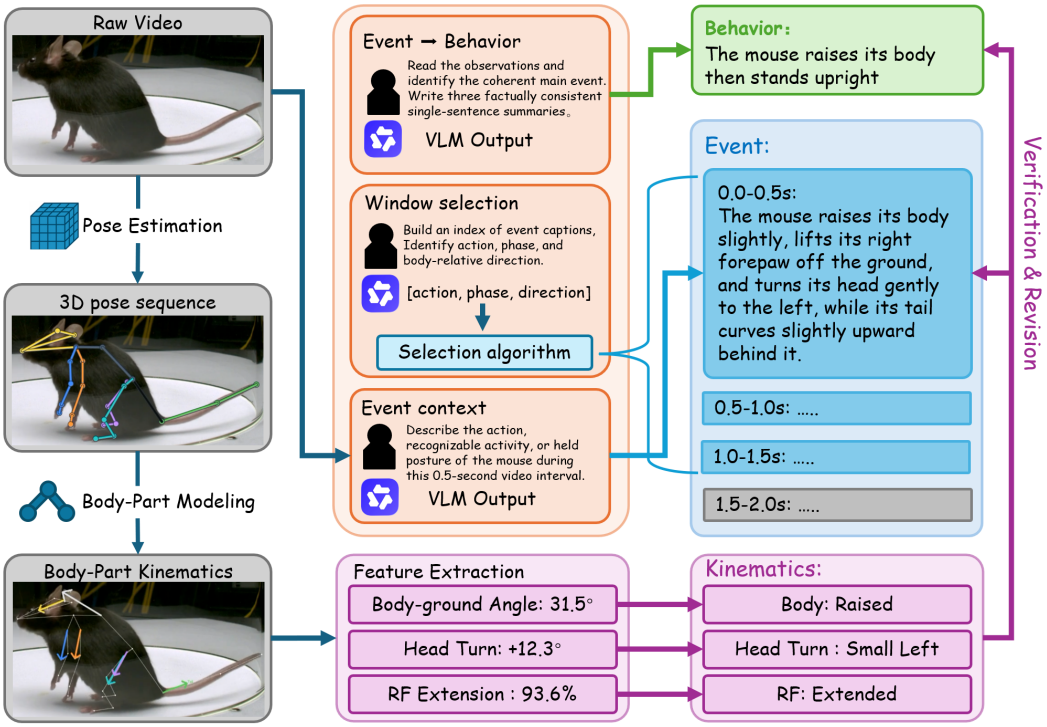

Figure 2: **RodentBEK dataset construction.** Pose-derived K supports verification and revision of video-derived E and event-based B summaries.

quantized motion as language across tasks (Jiang et al., 2023). We adapt the diffusion, autoregressive, and masked-token families to compare behavior and event conditioning on RodentBEK.

3 Dataset Construction

Figure 2 shows how continuous laboratory recordings become aligned motion–language pairs. Rodent25 reconstruction supplies trajectories (Section 3.1), from which we compute K (Section 3.2). Video-derived E guides window selection and B summaries, with K supporting verification and revision (Sections 3.3–3.4). We then partition and standardize the paired records (Section 3.5).

3.1 Source Recordings and Rodent25 Reconstruction

Our sources are Rat-7M (Dunn et al., 2021) and the multi-view mouse recordings of Li et al. (2023). We use five mouse recordings sampled at 100 Hz and seven rat recordings sampled at 120 Hz. Rodent25 defines corresponding landmarks across the two species: the snout and ears, three spine landmarks, three tail landmarks, and four landmarks along each limb.

We train DANNCE models with manual annotations in the Rodent25 representation, use s-DANNCE refinement, and apply additional corrections for difficult poses and tail reconstruction (Dunn et al., 2021; Klibaite et al., 2025). This yields 1,116,000 reconstructed mouse frames and 1,161,592 reconstructed rat frames, totaling 2,277,592 frames across the two sources (Figure 3).

Table 2 compares source coordinates with Rodent25 predictions screened to retain 99% of frames by confidence. Valid measures coordinate availability, Adj. consecutive valid observations, and the coefficient of variation (CV) native limb-length variation. Mouse CV approaches the sparse-label reference. Four of seven rat recordings have lower CV than source mocap, but s5d1 and s5d2 raise the equally weighted mean slightly above it. CV uses each resource's native limb definitions. Incomplete views and suspected calibration issues may contribute to these differences (Appendix C).

Table 2: **Landmark coverage and coordinate consistency.** Rodent25 Valid is a 99% QC retention setting; source Valid counts finite coordinates (†: sparse labels).

Species	Representation	Landmarks	Frames	Valid (%)	CV (%)	Adj. (%)
Mouse	Source labels	22	310†	96.51	14.41	—
	Rodent25	25	1,116,000	99.00	15.07	98.82
Rat	Source mocap	20	1,161,592	93.66	28.54	93.30
	Rodent25	25	1,161,592	99.00	29.43	98.76

On manual labels overlapping model development, mean per-joint position error (MPJPE) is 2.28 mm for mouse and 8.45 mm for rat. Generation uses the preprocessing masks in Appendix B.5.

3.2 Kinematic Evidence

K expresses posture and movement as measurements of seven parts: the body, head, four limbs, and tail.

Part directions. After filtering anomalous bone lengths, interpolating short internal gaps, and smoothing valid trajectories, we estimate a direction for each part. Let $\mathbf{p}_j(t)$ denote a landmark position and $\nu(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$ denote normalization of a valid nonzero vector. The part direction is

$$\mathbf{d}_b(t) = \begin{cases} \mathrm{orient}(\mathbf{v}_1(\mathbf{X}_b(t)),\, \mathbf{p}_{\mathrm{SpineF}}(t) - \mathbf{p}_{\mathrm{SpineL}}(t)), & b = \mathrm{Body}, \\ \nu(\mathbf{p}_{b,\mathrm{dist}}(t) - \mathbf{p}_{b,\mathrm{base}}(t)), & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathbf{X}_b$ contains the centered spine landmarks, $\mathbf{v}_1$ is their unit first principal component, and orient fixes its sign toward the front of the body. Head anchors are the ear midpoint and snout; limbs and tail use anatomical endpoints.

Relative motion. We express posture and movement in a reference frame as

$$\mathbf{u}_{b|r} = \mathbf{R}_r^\top \mathbf{d}_b, \qquad \mathbf{x}_{b|r} = \mathbf{R}_r^\top (\mathbf{p}_b - \mathbf{a}_r)/L, \qquad \Delta f = \mathrm{med}_{\mathrm{end}}\, f - \mathrm{med}_{\mathrm{start}}\, f. \tag{2}$$

Here $\mathbf{R}_r$ gives reference axes, $\mathbf{a}_r$ an anchor, and L the median spine length; time arguments are omitted. This form accommodates both ground and body references. Direction components determine angles; endpoint medians summarize angular or positional changes, while within-interval variation captures movements that return toward their starting state.

Internal extension. For a limb or tail chain $\mathbf{q}_0, \ldots, \mathbf{q}_n$, internal extension is measured by

$$\rho_b(t) = \frac{\|\mathbf{q}_n(t) - \mathbf{q}_0(t)\|_2}{\max\left\{\epsilon,\, \sum_{i=0}^{n-1} \|\mathbf{q}_{i+1}(t) - \mathbf{q}_i(t)\|_2\right\}}. \tag{3}$$

Here $\epsilon > 0$ prevents division by zero, and a larger ratio indicates a straighter chain. We aggregate these measurements within each 0.5 s interval and map them to language labels using feature-specific thresholds. For example, the raised-body state illustrated in Figure 2 corresponds to a trunk inclination of $15°$–$45°$ relative to the ground. K retains the measurements, language labels, and validity records used to verify descriptions.

3.3 Event Annotation and Event-Guided Windows

E records the visible action, local movement, or held posture within each 0.5 s interval. Qwen3.8-Flash (Alibaba Cloud, 2026) reads four sampled video timepoints per interval to generate an initial description, with visibility and uncertainty stored separately.

Alongside fixed four-second clips, we construct variable-length windows by combining consecutive events. The VLM converts E into an index of actions, phases, body-relative directions, and uncertainty (Figure 2). A deterministic algorithm then selects windows of 2, 2.5, 3, 3.5, or 4 seconds on

the half-second grid. The algorithm favors coherent processes and informative boundaries, selects at most one candidate per two-second anchor, and removes redundant overlaps. Each selected window reuses E and K from the corresponding source intervals, yielding four to eight aligned event slots.

3.4 Behavior Summaries and Evidence-Based Verification

B summarizes the ordered E descriptions within a window, for both fixed and event-guided clips. We generate three alternative captions with complementary emphases: the main actions, the recognizable activity and its essential movement, and a concise account of the process. Deriving B from the localized events keeps the two language levels tied to the same observed actions. Among four generation routes, event-based summarization with post-verification receives the highest mean Correctness, Coverage, and Controlled Diversity ratings (Table 10).

K supports the annotation model's verification of movement claims in B and E. A rule-based posture check also identifies conflicts between upright-posture descriptions and reconstructed trunk inclination. For example, a claim that the animal rises upright conflicts with reliable measurements that remain close to horizontal throughout the interval. Conflicting statements and the supporting K measurements are included in the prompt when regenerating a description. Visual context provides cues for activities such as grooming; K verifies their physical details.

Annotation uses Qwen3.8-Flash at temperature zero; prompts, responses, reliability rules, and revisions are retained for traceability.

3.5 Benchmark Preparation and Statistics

The construction archive contains 12,720 windows. Quality screening, deduplication, and temporal isolation yield the 8,195 retained windows in Figure 3. Outside designated test regions, 30-second source blocks define training and validation. A two-second margin on each side of a partition boundary provides at least four seconds of separation. Overlapping windows are permitted only within the training partition; no source frames are shared across partitions. The split separates source time, with animals and sessions potentially present in multiple partitions.

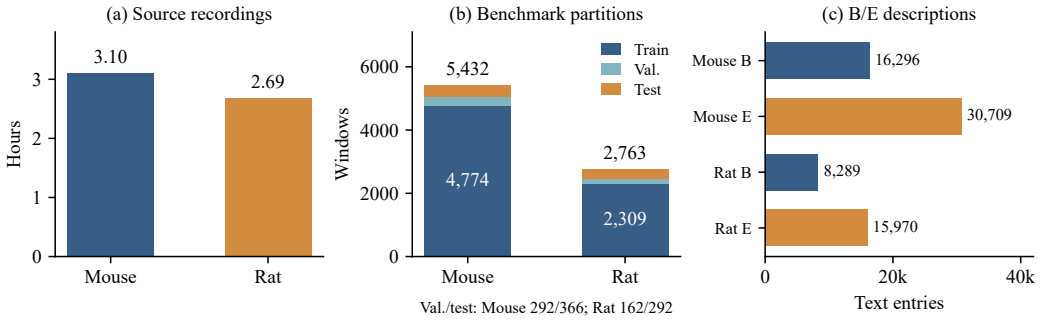

Figure 3: **RodentBEK at a glance.** Reconstructed source duration used by RodentBEK, retained windows by split, and B/E text counts.

Screening checks observation coverage, landmark validity, and root trajectories. Motion is standardized in meters at 30 Hz, retaining root displacement, species scale, true lengths, and reliability masks. Shared normalization uses pooled training data only; Appendix B.5 specifies gap handling and transforms.

Original empty event slots are represented by `<Motionless>`, preserving the no-motion annotation convention. Under this convention, 99.22% of mouse slots and 99.59% of rat slots in the construction archive contain motion descriptions, compared with 87.88% and 89.17% in the length-matched human and automatic FineMotion subsets (Table 11). Under the respective annotation pipelines, motion descriptions apply to a higher proportion of these rodent intervals at the shared 0.5-second resolution.

The benchmark has 7,083 training, 454 validation, and 658 test windows. Text counts comprise 24,585 B captions and 46,679 nonempty E descriptions, excluding 463 Motionless slots. Window and text counts include overlapping training coverage; synchronized timepoints are counted once across cameras (Table 14). Derived records retain source attribution, provenance, and redistribution terms.

4 EXPERIMENTS

We examine how E changes whole-window alignment, motion distributions, and local event correspondence. For each model and training seed, B and B+E use the same jointly trained mouse–rat checkpoint.

4.1 EXPERIMENTAL SETUP

Data and conditioning. We use the partitions in Section 3.5, including 366 mouse and 292 rat test windows. All models operate on the common Rodent25 representation at 30 Hz with explicit target lengths. Each model learns both B and B+E inputs during training. All three models separately encode and pool B and E with frozen T5-base, then concatenate the representations (Raffel et al., 2020); B-only input replaces E with <EMPTY>. Appendix A specifies training sampling and conditioning. Generation and the main semantic metrics use the first behavior caption, B0; supplementary multi-positive retrieval uses all three captions.

Baselines. We adapt MDM (Tevet et al., 2023), T2M-GPT (Zhang et al., 2023a), and MoMask (Guo et al., 2024). MDM generates continuous coordinates through diffusion; T2M-GPT generates tokens from a frozen single-level motion codec; MoMask combines residual quantization with masked and residual token prediction. B/B+E pairs share all generator and codec weights, selected by architecture-specific validation objectives (Appendix A).

Motion–text evaluators. We train separate mouse and rat evaluators on real training pairs and select checkpoints using real validation pairs. Each motion encoder takes root-relative landmark positions, root translation, velocities, and validity masks, scaled using training data. A two-layer Transformer with positional encoding maps these features to normalized 128-dimensional global and local embeddings; a learned text head projects frozen T5-base features (Raffel et al., 2020) into the same space. Following contrastive motion–language alignment (Petrovich et al., 2023), training combines a symmetric multi-positive global loss with a half-weighted local event loss:

$$\mathcal{L}_{\text{eval}} = \mathcal{L}_{\text{global}} + 0.5\mathcal{L}_{\text{event}}. \tag{4}$$

Motions paired with identical descriptions are treated as positives rather than negatives. Appendix A.2 gives training settings, loss definitions, and diagnostic checks.

Metrics. We evaluate semantic agreement, distribution similarity, and generated variation (Guo et al., 2022; Zhang et al., 2023a). R-Precision ranks B0 against 31 distinct negative descriptions for each motion; other valid B alternatives are excluded from the negatives. MM-Dist is the mean Euclidean distance between paired motion and B0 embeddings. FID compares feature means and covariances (Heusel et al., 2017) using our rodent evaluators. Diversity measures distances between different generated samples; its target is the GT value ($\rightarrow$). Multimodality measures distances between generations under the same condition. GT denotes reconstructed reference motion and uses the same semantic and diversity queries as the generators. Separate real–real (RR) distribution references, unbiased MMD^2 (Bińkowski et al., 2018), and multi-positive text-to-motion retrieval appear in the appendix.

Repeated evaluation. Each of three training seeds supplies 20 generated test sets. We evaluate them using 20 shared evaluation seeds, which resample retrieval candidates, diversity pairs, and RR block partitions. In round r, the three training-seed scores are averaged to m_r. We report $\bar{m} \pm t_{0.975,19} s_m/\sqrt{20}$, a 95% confidence interval conditional on the fitted models. GT uses the same sampling protocol; deterministic full-set scores have zero run variance. Training-seed variation is reported separately in Table 8.

4.2 BENCHMARK RESULTS

Semantic alignment and distribution similarity favor different models (Table 3). MoMask leads generated-motion retrieval for both species. For mouse, it also achieves the lowest FID, combining behavior-caption alignment with a closer aggregate feature distribution. For rat, T2M-GPT instead obtains the lowest FID while MoMask retains stronger retrieval. Retrieval distinguishes descriptions, whereas FID summarizes the generated set as a whole. Generated retrieval can exceed GT: generation explicitly targets B0, whereas real trajectories can contain valid variation left unspecified by that caption.

Table 3: Joint mouse–rat generation (mean and 95% CI, 20 rounds). Top-1/3: R-Precision; bold marks the best generated Top-1/3, FID, and MM-Dist per species.

Species	Model	Input	Top-1 ↑ (%)	Top-3 ↑ (%)	FID ↓	MM-Dist ↓	Diversity →
Mouse	GT	–	$27.27^{\pm0.98}$	$60.19^{\pm0.71}$	–	$1.099^{\pm0.000}$	$1.247^{\pm0.003}$
	MDM	B	$15.27^{\pm0.57}$	$36.63^{\pm0.73}$	$0.231^{\pm0.003}$	$1.184^{\pm0.002}$	$1.135^{\pm0.004}$
		B+E	$15.54^{\pm0.78}$	$36.96^{\pm0.85}$	$0.247^{\pm0.003}$	$1.180^{\pm0.002}$	$1.122^{\pm0.005}$
	T2M-GPT	B	$14.49^{\pm0.48}$	$35.67^{\pm0.93}$	$0.199^{\pm0.002}$	$1.179^{\pm0.002}$	$1.248^{\pm0.002}$
		B+E	$14.74^{\pm0.53}$	$35.72^{\pm0.89}$	$0.187^{\pm0.001}$	$1.175^{\pm0.002}$	$1.235^{\pm0.002}$
	MoMask	B	$\mathbf{28.03}^{\pm0.67}$	$\mathbf{60.39}^{\pm0.66}$	$0.138^{\pm0.001}$	$\mathbf{1.094}^{\pm0.001}$	$1.207^{\pm0.003}$
		B+E	$24.45^{\pm0.62}$	$54.45^{\pm0.66}$	$\mathbf{0.125}^{\pm0.001}$	$1.103^{\pm0.001}$	$1.200^{\pm0.003}$
Rat	GT	–	$9.52^{\pm0.52}$	$24.33^{\pm0.88}$	–	$1.290^{\pm0.000}$	$1.334^{\pm0.002}$
	MDM	B	$15.37^{\pm0.69}$	$34.61^{\pm0.56}$	$0.297^{\pm0.004}$	$1.269^{\pm0.001}$	$1.349^{\pm0.002}$
		B+E	$15.42^{\pm0.40}$	$35.48^{\pm0.55}$	$0.297^{\pm0.003}$	$1.263^{\pm0.002}$	$1.343^{\pm0.002}$
	T2M-GPT	B	$12.45^{\pm0.49}$	$29.81^{\pm0.70}$	$\mathbf{0.181}^{\pm0.002}$	$1.282^{\pm0.001}$	$1.339^{\pm0.002}$
		B+E	$12.49^{\pm0.52}$	$29.77^{\pm0.55}$	$0.195^{\pm0.002}$	$1.281^{\pm0.002}$	$1.340^{\pm0.002}$
	MoMask	B	$\mathbf{19.63}^{\pm0.45}$	$\mathbf{42.79}^{\pm0.60}$	$0.200^{\pm0.001}$	$\mathbf{1.234}^{\pm0.001}$	$1.323^{\pm0.001}$
		B+E	$17.37^{\pm0.47}$	$39.33^{\pm0.54}$	$0.190^{\pm0.001}$	$1.247^{\pm0.001}$	$1.331^{\pm0.001}$

For rat, Diversity remains close to the GT value across models, showing similar across-condition spread alongside differences in semantic alignment (Table 3).

4.3 EFFECT OF EVENT-LEVEL CONDITIONING

For MDM and T2M-GPT, adding E produces small increases in mean Top-1 retrieval: 0.27 and 0.25 percentage points for mouse, and less than 0.1 point for rat. Their lower MM-Dist is consistent with closer average alignment to B0, although the effect on Top-3 is mixed. The distributional response depends on the model and species: T2M-GPT improves mouse FID but increases rat FID, while MDM's mouse FID increases and its rat FID remains almost unchanged. These results separate closer caption alignment from changes in the aggregate motion distribution.

MoMask shows the complementary pattern. E reduces FID by 8.9% for mouse and 4.8% for rat, but lowers Top-1 by approximately 3.6 and 2.3 percentage points, respectively. This combination calls for direct evaluation of E correspondence alongside whole-window B alignment.

For MDM and T2M-GPT, mouse Multimodality decreases by 15.2% and 5.8%, respectively (Table 7). This is consistent with more specific descriptions narrowing the range of compatible motions.

4.4 LOCAL EVENT CORRESPONDENCE

Local evaluation tests whether generated intervals match their E descriptions. The frozen local heads distinguish correct pairs from mismatched E, shifted events, and static motion (Appendix A.3). Test evaluation covers 2,807 mouse and 2,053 rat intervals with valid E and pose. Event Recall@K (Event R@K) ranks the paired E among 32 descriptions, reported at $K = 1, 3$; Event MM-Dist measures paired embedding distance. Scores average intervals within windows, then windows.

Table 4: Local event correspondence, averaged first within windows (mean and 95% CI).

		Event R@1 (%) ↑		Event R@3 (%) ↑		Event MM-Dist ↓	
Species	Model	B	B+E	B	B+E	B	B+E
Mouse	MDM	$6.92^{\pm 0.21}$	$8.00^{\pm 0.12}$	$18.23^{\pm 0.39}$	$20.40^{\pm 0.27}$	$1.2863^{\pm 0.0008}$	$1.2735^{\pm 0.0007}$
	T2M-GPT	$8.14^{\pm 0.25}$	$8.94^{\pm 0.17}$	$21.14^{\pm 0.35}$	$22.98^{\pm 0.25}$	$1.2861^{\pm 0.0008}$	$1.2734^{\pm 0.0007}$
	MoMask	$11.72^{\pm 0.12}$	$15.07^{\pm 0.14}$	$28.52^{\pm 0.18}$	$34.52^{\pm 0.15}$	$1.2620^{\pm 0.0003}$	$1.2454^{\pm 0.0003}$
Rat	MDM	$6.20^{\pm 0.19}$	$7.21^{\pm 0.19}$	$16.51^{\pm 0.28}$	$18.76^{\pm 0.24}$	$1.3202^{\pm 0.0006}$	$1.3094^{\pm 0.0008}$
	T2M-GPT	$5.83^{\pm 0.14}$	$6.81^{\pm 0.17}$	$15.91^{\pm 0.20}$	$18.14^{\pm 0.27}$	$1.3035^{\pm 0.0006}$	$1.2958^{\pm 0.0005}$
	MoMask	$6.34^{\pm 0.13}$	$7.62^{\pm 0.19}$	$17.26^{\pm 0.22}$	$19.86^{\pm 0.14}$	$1.2970^{\pm 0.0002}$	$1.2896^{\pm 0.0002}$

E improves all local metrics across models and species (Table 4). MoMask's mouse Event R@1 gains 3.35 points even though its global Top-1 decreases under B+E, showing that closer local correspondence can coexist with weaker whole-window caption discrimination. MDM and T2M-GPT also gain more in local than global retrieval. Gains at both cutoffs and in paired distance support evaluating the added E descriptions directly, alongside whole-window B alignment.

5 Discussion

The results highlight three directions for extending the benchmark and its models.

Species and environments. Laboratory recordings support detailed reconstruction of mice and rats. Extending coverage requires wild settings and other rodents, including the hamsters, beavers, porcupines, and prairie dogs represented in existing resources (Table 12).

Behavior and interaction. Turning and upright posture are frequent in E; climbing/jumping appears in fewer than 0.2% of unique source-time slots (Table 13). All recordings contain a single animal. Future extensions could include richer environmental and rodent–rodent interactions, complementing paired-animal resources such as PAIR-R24M (Marshall et al., 2021).

Modeling and evaluation. The baselines pool E into one vector. Preserving event positions could strengthen interval-specific conditioning; part-aware codecs and anatomical constraints could better capture small movements. K could also constrain generated body-part states, extending its current role in annotation verification. Matched adaptation budgets and independent human ratings would extend model comparison. Since animals occur across source-time partitions, additional subjects and environments are needed to study generalization beyond the recorded conditions.

6 Conclusion

RodentBEK pairs 8,195 laboratory mouse and rat motion windows with behavior summaries, local events, and pose-derived kinematic evidence. Across three generator families, event conditioning improves local correspondence, while global semantic alignment and distribution similarity respond differently. These complementary measures support controlled study of fine-grained rodent motion generation.

REPRODUCIBILITY STATEMENT

Reproducibility details, including generator configurations, evaluator training, metric definitions, preprocessing rules, and repeated-evaluation protocols, are provided in Appendices A–C. Code and evaluation artifacts will be released with the benchmark.

AI USE STATEMENT

AI tools assisted literature discovery, manuscript editing, evaluation design, result interpretation, and the implementation of plotting and evaluation scripts. The annotation pipeline uses Qwen3.8-Flash as described in Section 3. All reported statistics were computed from recorded data and saved model outputs, with calculation scripts and evaluation records retained for reproducibility.

REFERENCES

Alibaba Cloud. Qwen3.8-Flash. Model Studio model documentation, 2026. URL `https://www.alibabacloud.com/help/en/model-studio/qwen3-8-flash`. Accessed September 26, 2026.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Sandeep Robert Datta, David J. Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. Computational neuroethology: A call to action. *Neuron*, 104(1):11–24, 2019. doi: 10.1016/j.neuron.2019.09.038.

Timothy W. Dunn, Jesse D. Marshall, Kyle S. Severson, Diego E. Aldarondo, David G. C. Hildebrand, Selmaan N. Chettih, William L. Wang, Amanda J. Gellis, David E. Carlson, Dmitriy Aronov, Winrich A. Freiwald, Fan Wang, and Bence P. Ölveczky. Geometric deep learning enables 3D kinematic profiling across species and environments. *Nature Methods*, 18(5):564–573, 2021. doi: 10.1038/s41592-021-01106-6.

Alex Gomez-Marin, Joseph J. Paton, Adam R. Kampff, Rui M. Costa, and Zachary F. Mainen. Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nature Neuroscience*, 17(11):1455–1462, 2014. doi: 10.1038/nn.3812.

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

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*, pp. 1900–1910, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alexander I. Hsu and Eric A. Yttri. B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature Communications*, 12(1):5188, 2021. doi: 10.1038/s41467-021-25420-x.

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*, volume 36, pp. 20067–20079, 2023.

Pierre Karashchuk, Katie L. Rupp, Evyn S. Dickinson, Sarah Walling-Bell, Elischa Sanders, Eiman Azim, Bingni W. Brunton, and John C. Tuthill. Anipose: A toolkit for robust markerless 3D pose estimation. *Cell Reports*, 36(13):109730, 2021. doi: 10.1016/j.celrep.2021.109730.

Ugne Klibaite, Tianqing Li, Diego Aldarondo, Jumana F. Akoad, Bence P. Ölveczky, and Timothy W. Dunn. Mapping the landscape of social behavior. *Cell*, 188(8):2249–2266.e23, 2025. doi: 10.1016/j.cell.2025.01.044.

Wonkwang Lee, Jongwon Jeong, Taehong Moon, Hyeon-Jong Kim, Jaehyeon Kim, Gunhee Kim, and Byeong-Uk Lee. How to move your dragon: Text-to-motion synthesis for large-vocabulary objects. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33110–33128. PMLR, 2025.

Tianqing Li, Kyle S. Severson, Fan Wang, and Timothy W. Dunn. Improved 3D markerless mouse pose estimation using temporal semi-supervision. *International Journal of Computer Vision*, 131 (6):1389–1405, 2023. doi: 10.1007/s11263-023-01756-3.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5442–5451, 2019.

Jesse D. Marshall, Ugne Klibaite, Amanda Gellis, Diego E. Aldarondo, Bence P. Ölveczky, and Timothy W. Dunn. The PAIR-R24M dataset for multi-animal 3D pose estimation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/1ff8a7b5dc7a7d1f0ed65aaa29c04b1e-Abstract-round1.html`.

Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 2018. doi: 10.1038/s41593-018-0209-y.

Arne Monsees, Kay-Michael Voit, Damian J. Wallace, Juergen Sawinski, Edyta Charyasz, Klaus Scheffler, Jakob H. Macke, and Jason N. D. Kerr. Estimation of skeletal kinematics in freely moving rodents. *Nature Methods*, 19(11):1500–1509, 2022. doi: 10.1038/s41592-022-01634-9.

Talmo D. Pereira, Joshua W. Shaevitz, and Mala Murthy. Quantifying behavior to understand the brain. *Nature Neuroscience*, 23(12):1537–1549, 2020. doi: 10.1038/s41593-020-00734-z.

Talmo D. Pereira, Nathaniel Tabris, Arie Matsliah, David M. Turner, Junyu Li, Shruthi Ravindranath, Eleni S. Papadoyannis, Edna Normand, David S. Deutsch, Z. Yan Wang, Grace C. McKenzie-Smith, Catalin C. Mitelut, Marielisa Diez Castro, John D'Uva, Mikhail Kislin, Dan H. Sanes, Sarah D. Kocher, Samuel S.-H. Wang, Annegret L. Falkner, Joshua W. Shaevitz, and Mala Murthy. SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods*, 19(4): 486–495, 2022. doi: 10.1038/s41592-022-01426-1.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9488–9497, 2023.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. doi: 10.1089/big.2016.0028.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with English labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 722–731, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

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

Xuan Wang, Kai Ruan, Xing Zhang, and Gaoang Wang. AniMo: Species-aware model for text-driven animal motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1929–1939, 2025.

Caleb Weinreb, Jonah E. Pearl, Sherry Lin, Mohammed Abdal Monium Osman, Libby Zhang, Sidharth Annapragada, Eli Conlin, Red Hoffmann, Sofia Makowska, Winthrop F. Gillis, Maya Jay, Shaokai Ye, Alexander Mathis, Mackenzie W. Mathis, Talmo Pereira, Scott W. Linderman, and Sandeep Robert Datta. Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics. *Nature Methods*, 21(7):1329–1339, 2024. doi: 10.1038/s41592-024-02318-2.

Alexander B. Wiltschko, Tatsuya Tsukahara, Ayman Zeine, Rockwell Anyoha, Winthrop F. Gillis, Jeffrey E. Markowitz, Ralph E. Peterson, Jesse Katon, Matthew J. Johnson, and Sandeep Robert Datta. Revealing the structure of pharmacobehavioral space through motion sequencing. *Nature Neuroscience*, 23(11):1433–1443, 2020. doi: 10.1038/s41593-020-00706-3.

Bizhu Wu, Jinheng Xie, Meidan Ding, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. FineMotion: A dataset and benchmark with both spatial and temporal annotation for fine-grained motion generation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13837–13846, 2025.

Zhangsihao Yang, Mingyuan Zhou, Mengyi Shan, Bingbing Wen, Ziwei Xuan, Mitch Hill, Junjie Bai, Guo-Jun Qi, and Yalin Wang. OmniMotionGPT: Animal motion generation with limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1249–1259, 2024.

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.

Zeyu Zhang, Yiran Wang, Biao Wu, Shuo Chen, Zhiyuan Zhang, Shiya Huang, Wenbo Zhang, Meng Fang, Ling Chen, and Yang Zhao. Motion avatar: Generate human and animal avatars with arbitrary motion. In *British Machine Vision Conference*, 2024.

Brian Nlong Zhao, Zhuoyang Pan, James M. Rehg, Jiajun Wu, and Shangzhe Wu. Kirin: Animal motion generation from in-the-wild video. In *Computer Vision – ECCV 2026*, volume 17055 of *Lecture Notes in Computer Science*, pp. 1–19. Springer, 2026. doi: 10.1007/978-3-032-37261-1_1.

A Implementation and Evaluation Details

A.1 Generator Configurations

MDM and T2M-GPT sample one B caption per window and enable E with probability 0.5. Source-time coverage weights reduce repeated exposure to overlapping windows; effective batches contain non-overlapping intervals. For these two models, B and the complete E script are separately encoded by frozen T5-base (Raffel et al., 2020), mean-pooled with attention masks into two 768-dimensional vectors, and concatenated. The B setting replaces the E script with <EMPTY>.

MDM uses 50 cosine diffusion steps, hidden width 512, eight Transformer layers, EMA decay 0.9999, learning rate 2×10^{-4}, and effective batch size 32. Validation occurs every 1,000 steps, with early stopping eligible after 20,000 steps; ten validations without 0.5% relative improvement trigger stopping, subject to a 100,000-step ceiling. The validation-selected EMA checkpoints for seeds 0, 1, and 2 occur at 53,000, 78,000, and 69,000 steps. Inference uses guidance 2.5.

T2M-GPT uses a frozen 512-entry XYZ75 codec with temporal downsampling by four, shared across its three generator seeds. The generator has width 1,024 and nine layers, learning rate 2×10^{-4}, effective batch size 32, and token retention probability 0.5. Clean-token validation occurs every 500 steps and ranks all saved checkpoints. Early stopping is eligible after 10,000 steps with patience ten, a 0.5% improvement threshold, and a 100,000-step ceiling. The best validation checkpoints occur at 8,000, 7,000, and 7,500 steps, respectively; inference temperature is 1.0. For joint mouse–rat training, MDM and T2M-GPT validation objectives average equally across species and across B and B+E. Initialization and data sampling use seeds $s \in \{0, 1, 2\}$; condition sampling uses $10{,}000 + s$, and validation uses seed 177.

MoMask uses its own six-level residual quantization codec with 512-entry codebooks and code width 512. Its adapter also concatenates separately pooled 768-dimensional T5 features for B and the complete E script. The resulting 1,536-dimensional condition is supplied to both the masked and residual token generators. B-only evaluation replaces the E vector with the encoding of <EMPTY>; training samples B alternatives and enables E with probability 0.5. Its generation network has latent width 384, eight layers, six attention heads, and feed-forward width 1,024. MoMask uses batch size 2,048, learning rate 2×10^{-4}, and stage budgets of 100,000 steps, with validation every 500 steps, earliest stopping at 10,000 steps, and patience ten. Generation uses ten mask iterations, masked guidance 4.0, residual guidance 2.0, temperature 1.0, and top-k ratio 0.9. For every training seed, B and B+E use identical masked, residual, and codec checkpoints when evaluated on either species. These architecture-specific configurations share the benchmark data and evaluation protocol.

A.2 Evaluator Training and Metric Definitions

The evaluator is trained independently for each species and frozen before scoring generators. Its motion input concatenates 25 root-relative landmark positions, root translation, their velocities, and the corresponding validity indicators (312 channels). The 156 position/velocity channels are divided by training-set root-mean-square scales. A linear projection feeds two Transformer layers with width 128, four attention heads, feed-forward width 512, dropout 0.1, and sinusoidal positions. The global head combines valid-frame mean features with the last valid frame; the local head pools within individual 15-frame intervals whose attention is evaluated separately. Both motion heads and the two-layer T5 feature projection output unit-normalized 128-dimensional embeddings. Text features come from frozen T5-base (Raffel et al., 2020). Root-relative coordinates are valid only where both the landmark and SpineM coordinates are valid. Velocities use consecutive valid coordinates at 30 Hz; the first velocity in each encoded window or local interval is masked. Invalid signal channels are zeroed after scaling. Attention and pooling exclude frames beyond the true length or without valid position channels; coordinate-validity indicators remain explicit inputs.

Training draws batches of 32 non-overlapping source intervals using source-coverage weights and samples one of the three B captions per window. AdamW uses learning rate 10^{-4} and gradient-norm clipping at 1. The global objective uses symmetric multi-positive contrastive learning, following motion–text alignment in Petrovich et al. (2023). For motion embeddings $\mathbf{z}_i$, text embeddings $\mathbf{y}_j$, positive set $P(i)$ defined by identical descriptions, and learned temperature τ, its motion-to-text

term is

$$\ell_{m\to t} = -\frac{1}{N}\sum_i \log \frac{\sum_{j\in P(i)} \exp(\mathbf{z}_i^\top \mathbf{y}_j/\tau)}{\sum_j \exp(\mathbf{z}_i^\top \mathbf{y}_j/\tau)}, \qquad \mathcal{L}_{\text{global}} = \tfrac{1}{2}(\ell_{m\to t} + \ell_{t\to m}). \tag{5}$$

The local loss has the same form over valid E intervals and weight 0.5. The learned inverse temperature starts at $1/0.07$ and is capped at 100. Training uses seed 0 with a 20,000-step budget and early stopping after six non-improving validations once 5,000 steps have elapsed. Validation is run every 500 steps; the selected checkpoint maximizes validation retrieval among candidates passing the prescribed global checks. These checks compare real motion with noise, mask-only and static controls, randomized pairings, and within-recording/length subsets. Validation selects the mouse and rat checkpoints at 5,000 and 2,000 steps, respectively. The frozen weights, data fingerprints, and evaluation source hashes are retained with the evaluation artifacts.

Retrieval and paired agreement. The main table follows the single-positive motion-to-text convention of Zhang et al. (2023a): each motion query ranks its conditioning caption B0 among 32 distinct descriptions. The other 31 descriptions are sampled without replacement from the species' distinct test-caption pool, excluding all three valid B alternatives. The candidate random generator uses evaluation seed plus 10^6, separately from the supplementary protocol, and candidates are shared across models, conditioning settings, and GT. R-Precision Top-K is the fraction of motion queries whose B0 occurs among the first K candidates.

Supplementary text-to-motion retrieval includes every motion with the query's exact B caption as a positive and samples negatives to reach up to 32 candidates. We report Recall@K, the fraction of queries with at least one positive among the top K, and mean reciprocal rank (MRR), the mean inverse rank of the first positive. This follows motion retrieval terminology (Petrovich et al., 2023); exact-text duplicates are positives in our multi-caption setting. Scores within 10^{-10} of the highest positive score are tied; we compute the expected metric under uniformly random ordering of those ties. Each query has equal weight; the text-to-motion query set is deduplicated by exact caption, whereas main-table motion-to-text retrieval averages over motion windows. MM-Dist is $N^{-1}\sum_i \|\mathbf{z}_i - \mathbf{y}_{i,0}\|_2$. The supplementary three-caption cosine is $(3N)^{-1}\sum_i \sum_{j=0}^{2} \mathbf{z}_i^\top \mathbf{y}_{i,j}$, where the sum over i spans all windows.

Distribution measures. Following the Fréchet feature-distance construction (Heusel et al., 2017),

$$\text{FID} = \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_r\|_2^2 + \text{tr}\Big(\mathbf{C}_g + \mathbf{C}_r - 2(\mathbf{C}_r^{1/2}\mathbf{C}_g\mathbf{C}_r^{1/2})^{1/2}\Big). \tag{6}$$

FID is computed in the learned rodent motion space, with sample covariances using the $n-1$ denominator. We also use the unbiased MMD estimator (Bińkowski et al., 2018) with $k(\mathbf{a},\mathbf{b}) = (\mathbf{a}^\top\mathbf{b}/128+1)^3$:

$$\widehat{\text{MMD}}^2 = \frac{\sum_{i\neq j} k(\mathbf{g}_i,\mathbf{g}_j)}{n(n-1)} + \frac{\sum_{i\neq j} k(\mathbf{r}_i,\mathbf{r}_j)}{m(m-1)} - \frac{2\sum_{i,j} k(\mathbf{g}_i,\mathbf{r}_j)}{nm}. \tag{7}$$

Negative finite-sample estimates are retained; tables scale this statistic by 10^3. Generated distributions use the full species test set. RR compares two disjoint sets of source-time blocks, with 173 mouse or 139 rat windows per side in every round. Table 9 reports RR separately because its sample size and block composition differ from the full-test-set comparison used for generated motion.

Variation. Diversity is the mean feature distance over 2,000 sampled index pairs after removing self-pairs. Multimodality averages the same distance over ten distinct unordered generation pairs for each fixed prompt, selected from its 20 saved generations. The same sampling rules are used across models and conditions. For sampled distinct window pairs $\mathcal{P}$ and distinct draw pairs $\mathcal{Q}$, the measures are

$$\text{Div} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \|\mathbf{z}_i - \mathbf{z}_j\|_2, \qquad \text{MM} = \frac{1}{N|\mathcal{Q}|}\sum_i \sum_{(a,b)\in\mathcal{Q}} \|\mathbf{z}_i^{(a)} - \mathbf{z}_i^{(b)}\|_2. \tag{8}$$

GT has one trajectory per prompt, so its Multimodality is undefined.

Repeated evaluation and uncertainty. Evaluation seeds 260926–260945 independently resample candidates, index pairs, generation pairs, and RR block partitions. Round r uses saved generation draw r for each training seed s. All generators use the seed schedule $600000 + 10000r + 40\ell + o + c$, where ℓ is target frame count, o the offset within a length-grouped batch of 16, and c is zero for mouse or 10^7 for rat. Thus the 20 draws use distinct generation seeds, shared across B/B+E; training seeds control the fitted weights. The main result is based on $m_r = \frac{1}{3}\sum_{s=0}^{2} m_{s,r}$ and $\bar{m} = \frac{1}{20}\sum_{r=0}^{19} m_r$, and is reported as

$$\bar{m} \pm t_{0.975,19}\frac{s_m}{\sqrt{20}}, \qquad s_m^2 = \frac{1}{19}\sum_{r=0}^{19}(m_r - \bar{m})^2. \tag{9}$$

This interval describes generation and evaluation randomness conditional on the test set and fitted models. Table 8 separately reports sample SD across the three training-seed means, each averaged over 20 rounds; neither summary estimates variation across new animals. Deterministic full-set GT MM-Dist and cosine scores have zero evaluation-run variance. GT retrieval, Diversity, and RR vary with their sampled candidates or comparisons. All metrics use the same frozen species-specific evaluator weights across generators and conditioning settings.

Diagnostic perturbations. Table 5 compares reference motion with within-window frame shuffling, repetition of the first pose, and independent Gaussian coordinate noise of 5 mm standard deviation. Shuffling moves coordinate masks with frames; all controls preserve the original true lengths. Correct reference motion has lower MM-Dist than these controls for both species. Cyclically shifting caption assignments also increases MM-Dist. These perturbations test sensitivity to motion and pairing changes; local event checks follow below.

Table 5: Evaluator responses to motion and caption perturbations; lower MM-Dist indicates closer paired agreement.

Species	Input	MM-Dist ↓	3B cosine ↑
Mouse	Reference	1.099	0.377
	Shuffled frames	1.283	0.166
	Static pose	1.216	0.221
	5 mm noise	1.445	-0.041
	Mismatched text	1.194	0.261
Rat	Reference	1.290	0.139
	Shuffled frames	1.357	0.066
	Static pose	1.319	0.092
	5 mm noise	1.386	0.032
	Mismatched text	1.326	0.101

A.3 Local Event Evaluation

Local training and evaluation use the benchmark E annotations, aligned by sample ID and source interval. Each description is encoded with frozen T5-base, attention-mask mean pooling, and the learned text projection.

Validation controls. We first test the local representations on real validation motion. An eligible slot has nonempty E, a complete 15-frame interval within the true window length, and valid pose channels. The local encoder retains its training masks and computes attention separately within each slot. We compare each correct motion–E pair with a different-text description from another recording, a different-text slot in the same window, and a static repetition of the slot's first pose. The static control repeats the coordinate mask with the pose and uses descriptions mentioning movement, turning, raising, lowering, stepping, rotation, walking, or grooming. Temporal mismatches select

the next available different-text slot cyclically within the window. Exact text duplicates are never counted as mismatches.

For each control, we calculate the increase in paired embedding distance relative to the correct pair. We average differences within source-time blocks, then bootstrap block means 2,000 times using seed 260926 to obtain percentile 95% intervals. Before scoring generators, we require positive lower bounds for the cross-recording mismatch and static controls; within-window mismatch is checked separately. Both species pass all three checks (Table 6). The controls validate automatic feature sensitivity; independent human assessment of generated events remains a further evaluation direction.

Table 6: Local evaluator validation: control minus correct-pair distance, averaged by source-time block, with bootstrap 95% intervals.

Species	Control	Slots	Blocks	Distance increase [95% interval]
Mouse	Other-recording E	1,793	46	0.162 [0.150, 0.176]
	Within-window E shift	1,793	46	0.054 [0.044, 0.063]
	Static motion	1,033	46	0.078 [0.068, 0.087]
Rat	Other-recording E	1,022	37	0.097 [0.083, 0.113]
	Within-window E shift	1,022	37	0.020 [0.012, 0.028]
	Static motion	620	37	0.020 [0.014, 0.027]

Generation metrics. Eligibility retains 2,807 of 2,826 mouse slots and 2,053 of 2,126 rat slots, excluding empty labels and invalid intervals. For eligible test slots, Event R@K ranks the paired E among 32 distinct descriptions sampled from the species' test events: one matching text and 31 other exact-text-distinct labels. All generators and conditions share candidates in each of the 20 evaluation rounds; ties use expected rank under random tie ordering. Event MM-Dist averages Euclidean distances between paired local motion and E embeddings. Table 4 averages eligible slots within each window, then averages windows, giving short and long windows equal weight. Slot-weighted aggregation preserves the qualitative result: B+E improves both retrieval and paired distance for every model and species. All generated conditions share reference masks and true lengths; both averages exclude absent labels and padding. For each round, we average the three training-seed metrics and report the same Student-t interval as for global evaluation. GT uses the same candidates and masks; its fixed-pair distance has zero round variance.

A.4 Supplementary Results

Table 7 complements the joint mouse–rat comparison with text-to-motion retrieval, MMD^2, and Multimodality. Table 8 reports variation across training seeds, and Table 9 gives the real–real distribution reference.

Table 7: Text-to-motion retrieval, distribution, and variation measures (mean and 95% CI; MMD^2 scaled by 10^3).

Species	Model	Input	R@1 (%)	R@2 (%)	R@3 (%)	MMD^2 ↓ ($\times 10^3$)	Multimodality
Mouse	GT	–	$26.76^{\pm 0.43}$	$45.62^{\pm 0.50}$	$59.00^{\pm 0.45}$	–	–
	MDM	B	$13.73^{\pm 0.43}$	$25.02^{\pm 0.43}$	$34.56^{\pm 0.48}$	$2.208^{\pm 0.036}$	$0.610^{\pm 0.001}$
		B+E	$15.20^{\pm 0.32}$	$26.93^{\pm 0.55}$	$36.67^{\pm 0.53}$	$2.347^{\pm 0.029}$	$0.517^{\pm 0.001}$
	T2M-GPT	B	$13.65^{\pm 0.35}$	$24.41^{\pm 0.33}$	$33.26^{\pm 0.48}$	$0.776^{\pm 0.024}$	$1.009^{\pm 0.001}$
		B+E	$14.03^{\pm 0.30}$	$25.24^{\pm 0.38}$	$34.71^{\pm 0.37}$	$0.680^{\pm 0.014}$	$0.950^{\pm 0.001}$
	MoMask	B	$\mathbf{24.69}^{\pm 0.34}$	$\mathbf{41.41}^{\pm 0.37}$	$\mathbf{53.41}^{\pm 0.34}$	$0.433^{\pm 0.007}$	$0.526^{\pm 0.000}$
		B+E	$24.51^{\pm 0.31}$	$40.61^{\pm 0.32}$	$52.54^{\pm 0.38}$	$\mathbf{0.315}^{\pm 0.006}$	$0.515^{\pm 0.000}$
Rat	GT	–	$8.78^{\pm 0.25}$	$15.64^{\pm 0.29}$	$21.74^{\pm 0.32}$	–	–
	MDM	B	$8.92^{\pm 0.32}$	$16.60^{\pm 0.35}$	$23.33^{\pm 0.37}$	$1.995^{\pm 0.051}$	$0.851^{\pm 0.002}$
		B+E	$9.45^{\pm 0.27}$	$17.65^{\pm 0.39}$	$24.69^{\pm 0.44}$	$1.965^{\pm 0.039}$	$0.770^{\pm 0.002}$
	T2M-GPT	B	$7.88^{\pm 0.28}$	$14.43^{\pm 0.32}$	$20.47^{\pm 0.34}$	$\mathbf{0.809}^{\pm 0.020}$	$0.928^{\pm 0.001}$
		B+E	$8.32^{\pm 0.25}$	$15.37^{\pm 0.32}$	$21.87^{\pm 0.31}$	$0.942^{\pm 0.017}$	$0.907^{\pm 0.001}$
	MoMask	B	$\mathbf{10.95}^{\pm 0.30}$	$\mathbf{19.59}^{\pm 0.31}$	$\mathbf{26.99}^{\pm 0.28}$	$1.358^{\pm 0.012}$	$0.485^{\pm 0.001}$
		B+E	$10.88^{\pm 0.23}$	$19.47^{\pm 0.30}$	$26.80^{\pm 0.34}$	$1.107^{\pm 0.013}$	$0.511^{\pm 0.001}$

Table 8: Training variation: mean and sample SD of the three training-seed means under the single-positive protocol.

Species	Model	Input	Top-1 ↑ (%)	Top-3 ↑ (%)	FID ↓	MM-Dist ↓	Diversity →
Mouse	MDM	B	$15.27^{\pm 1.09}$	$36.63^{\pm 2.13}$	$0.231^{\pm 0.027}$	$1.184^{\pm 0.010}$	$1.135^{\pm 0.008}$
		B+E	$15.54^{\pm 1.16}$	$36.96^{\pm 2.66}$	$0.247^{\pm 0.035}$	$1.180^{\pm 0.014}$	$1.122^{\pm 0.010}$
	T2M-GPT	B	$14.49^{\pm 0.37}$	$35.67^{\pm 0.18}$	$0.199^{\pm 0.008}$	$1.179^{\pm 0.002}$	$1.248^{\pm 0.007}$
		B+E	$14.74^{\pm 0.50}$	$35.72^{\pm 0.82}$	$0.187^{\pm 0.008}$	$1.175^{\pm 0.003}$	$1.235^{\pm 0.018}$
	MoMask	B	$28.03^{\pm 0.79}$	$60.39^{\pm 1.56}$	$0.138^{\pm 0.007}$	$1.094^{\pm 0.004}$	$1.207^{\pm 0.004}$
		B+E	$24.45^{\pm 0.40}$	$54.45^{\pm 0.48}$	$0.125^{\pm 0.008}$	$1.103^{\pm 0.003}$	$1.200^{\pm 0.005}$
Rat	MDM	B	$15.37^{\pm 0.94}$	$34.61^{\pm 1.65}$	$0.297^{\pm 0.022}$	$1.269^{\pm 0.009}$	$1.349^{\pm 0.004}$
		B+E	$15.42^{\pm 0.97}$	$35.48^{\pm 1.56}$	$0.297^{\pm 0.018}$	$1.263^{\pm 0.008}$	$1.343^{\pm 0.004}$
	T2M-GPT	B	$12.45^{\pm 1.11}$	$29.81^{\pm 2.14}$	$0.181^{\pm 0.011}$	$1.282^{\pm 0.008}$	$1.339^{\pm 0.004}$
		B+E	$12.49^{\pm 0.78}$	$29.77^{\pm 1.21}$	$0.195^{\pm 0.013}$	$1.281^{\pm 0.003}$	$1.340^{\pm 0.006}$
	MoMask	B	$19.63^{\pm 0.44}$	$42.79^{\pm 0.45}$	$0.200^{\pm 0.011}$	$1.234^{\pm 0.002}$	$1.323^{\pm 0.010}$
		B+E	$17.37^{\pm 0.47}$	$39.33^{\pm 1.89}$	$0.190^{\pm 0.010}$	$1.247^{\pm 0.004}$	$1.331^{\pm 0.001}$

Table 9: Real–real distribution references from disjoint source-time blocks (mean and 95% CI).

Species	Windows per side	FID	MMD^2 ($\times 10^3$)
Mouse	173	$0.141^{\pm 0.011}$	$0.413^{\pm 0.174}$
Rat	139	$0.194^{\pm 0.008}$	$0.219^{\pm 0.106}$

B Dataset Comparisons and Event Statistics

B.1 Behavior Caption Generation

We compare event-to-behavior and direct video-to-behavior generation, each with and without post-verification. Each of three raters scores 200 caption sets per method on each criterion using a five-point scale; Table 10 averages the 600 ratings for each entry. Each rater views the video and assesses its three B captions as a set, with the generation method hidden (Figure 4). *Correctness* measures whether the described actions actually occur in the video. *Coverage* measures how completely the captions cover the observed actions. *Controlled Diversity* measures whether the captions offer varied wording and complementary descriptions while remaining factually consistent with one another. Higher scores indicate better performance.

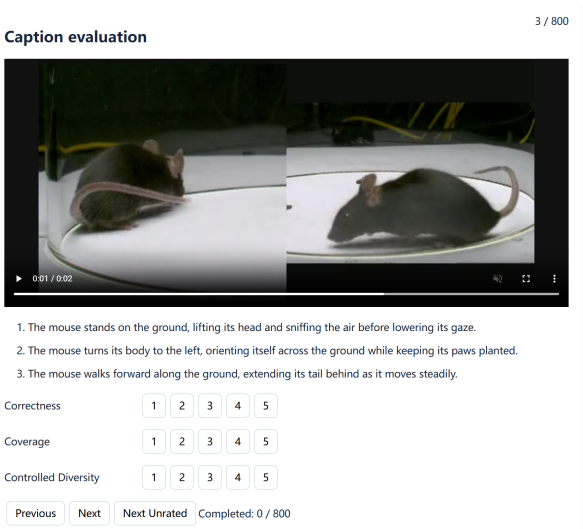

Figure 4: **Caption-rating interface.** Correctness, Coverage, and Controlled Diversity are rated with the generation method hidden.

Event-to-behavior generation with post-verification has the highest mean on all three criteria. Without post-verification, event-based generation also yields higher mean scores than direct video generation on all three criteria, supporting the use of local event descriptions to construct B.

Table 10: Behavior-caption ratings: mean Correctness, Coverage, and Controlled Diversity on a 1–5 scale.

ID	B generation route	Correctness	Coverage	Controlled Diversity
A	$E \to B$ + post-verification	**3.66**	**3.52**	**3.21**
B	$E \to B$	3.44	3.44	3.13
C	Video $\to B$ + post-verification	3.42	3.49	3.02
D	Video $\to B$	3.37	3.33	2.83

B.2 Event Density

We compare original RodentBEK E with the human and automatic short-interval descriptions released by FineMotion (Wu et al., 2025). Both locate descriptions at 0.5-second intervals. An empty original description corresponds to `<Motionless>`. Density is the fraction of actual, unpadded slots with nonempty descriptions. Word counts average over nonempty descriptions. Table 11 uses original E from the construction archive: 7,787 mouse and 4,933 rat windows, before curation into

the final 8,195-window benchmark. The comparison also includes FineMotion windows with four to eight slots to match the 2–4-second duration range.

Table 11: Motion-described interval density at 0.5-second resolution, including length-matched FineMotion subsets.

Data view	Windows	Slots	Nonempty	Density (%)	Mean words per E
FineMotion human, all	1,500	21,344	18,768	87.93	26.31
FineMotion auto, all	29,230	420,968	370,222	87.95	20.36
FineMotion human, 2–4 s	274	1,798	1,580	87.88	26.80
FineMotion auto, 2–4 s	4,968	32,470	28,952	89.17	19.72
Mouse, all	7,787	42,892	42,557	99.22	13.68
Mouse, fixed	2,790	22,320	22,150	99.24	13.70
Mouse, event-guided	4,997	20,572	20,407	99.20	13.66
Rat, all	4,933	28,091	27,977	99.59	14.84
Rat, fixed	2,089	16,712	16,651	99.64	14.82
Rat, event-guided	2,844	11,379	11,326	99.53	14.86

FineMotion human and auto densities are 87.93% and 87.95% for the full releases, and 87.88% and 89.17% in the length-matched subsets. RodentBEK original E reaches 99.22% for mouse and 99.59% for rat. At the shared 0.5-second resolution, both rodent sources, under their respective annotation pipelines, have a higher proportion of motion-described intervals than the compared FineMotion subsets. Overlapping windows reuse original E; counting each absolute source slot once yields 99.24% for mouse (22,320 slots) and 99.64% for rat (16,712 slots). The contrast persists after source-time deduplication, supporting the use of local descriptions for these recordings.

B.3 Sources, Species, and Scale

Table 12 details the sources, species, scale, and language organization of the resources in Table 1. Asset categories retain their release names; NR denotes unreported finer identification. The asset-based datasets broaden taxonomic coverage, while laboratory sources provide synchronized observations of actual animals. Rat-7M's source study identifies adult Long–Evans rats (Charles River 006); its developmental pup experiments are separate from the recordings used here (Dunn et al., 2021). The mouse source provides five recordings without recording-specific strain assignments (Li et al., 2023).

Zoo-300K's public rodent inventory contains 1,840 hamster, 2,220 rat, and 4,680 Sand Mouse clips; captions are released for hamster and Sand Mouse only. AniMo4D contributes 437 American beaver, 471 crested porcupine, and 377 prairie dog clips, with 1,021, 1,113, and 901 descriptions, respectively.

T2M4LVO's rodent inventory comprises 5 hamster, 7 rat, and 13 Sand Mouse clips. Its 75 base captions count each short/mid/long description once, separately from paraphrases and skeletal augmentation. Table 1 also lists their seven original-style subject variants (525 entries), which supply the Words/text average. Across species, the released base-caption inventory covers 1,135 clips and 3,405 descriptions; the paper reports 1,137 source animations.

Words/text uses released text without boilerplate removal, treating hyphenated words and contractions as single tokens. RodentBEK averages over each B or nonempty E entry, including text reused in overlapping windows. For Rat-7M, camera-frame counts and elapsed time are distinct: the audited public videos span 2.71 hours, while the reconstructed trajectories used here span 2.69 hours.

B.4 Behavior-Related Content of E

We identify each released half-second interval by species, recording, and absolute native-frame boundaries. This yields 18,840 unique mouse intervals and 11,678 rat intervals; repeated win-

Table 12: Public motion resources: (a) sources and rodent identities; (b) scale and language organization.

(a) Sources and species

Resource	Motion source / coverage	Rodent identities
AnimalML3D (Yang et al., 2024)	DeformingThings4D animation; 36 animal assets	None in the released inventory
Zoo-300K (Zhang et al., 2024)	Truebones animation; 65 source categories	Hamster, Rat, Sand Mouse
T2M4LVO (Lee et al., 2025)	Truebones animation with skeleton augmentation	Hamster, Rat, Sand Mouse
AniMo4D (Wang et al., 2025)	Planet Zoo animation; 114 species	American beaver, crested porcupine, prairie dog
Rat-7M (Dunn et al., 2021)	Laboratory multi-view video and marker capture	Laboratory rat; Long–Evans (Charles River 006)
3D Mouse Pose (Li et al., 2023)	Laboratory multi-view video; five mouse recordings	Laboratory mouse; recording-specific strain assignment NR
RodentBEK	Video-derived Rodent25 reconstruction; mouse and rat	Mice from Li et al.; Long–Evans rats from Rat-7M

(b) Scale and language organization

Resource	Original / augmented scale	Rodent subset	Language organization
AnimalML3D	1,240 clips; 3,720 reported captions	0	Three clip captions
Zoo-300K	878 source clips; approximately 300K augmented pairs reported	24 source clips; 8,740 released clips, 6,520 texts	Clip-level paragraphs
T2M4LVO	1,137 source clips reported; approximately 26K skeleton variants	25 clips; 75 base captions	Short/mid/long descriptions
AniMo4D	78,149 clips; 185,435 texts	1,285 clips; 3,035 texts	Clip-level descriptions
Rat-7M	Nearly 7M camera frames reported; 2.71 h in the audited public video release	Seven released recordings; 1,161,592 reconstructed timepoints used (2.69 h)	No paired generation captions
3D Mouse Pose	Five recordings; 3.10 h at 100 Hz	1,116,000 reconstructed timepoints used	No paired generation captions
RodentBEK	2,277,592 reconstructed timepoints; 12,720 construction windows	8,195 retained clips; 71,264 text entries	B/E hierarchy and K evidence

dows agree on their event text. Table 13 counts case-insensitive matches to a fixed vocabulary of movement and posture terms. Multiple matches per slot are allowed. The word families cover walking/running/stepping/crawling, turning/pivoting/rotation, upright or hind-leg posture, grooming/scratching/licking/washing, sniffing, stillness, eating/drinking, and climbing/jumping/leaping. These counts measure overlapping word mentions, including negated statements, rather than mutually exclusive behavior classes. The complete matching expressions and manifest hash accompany the statistical record. Turning is mentioned in 23.20% of mouse slots and 16.51% of rat slots; upright posture appears in 18.22% and 12.23%, respectively. Climbing or jumping is mentioned in only 3 mouse and 12 rat slots. The scarcity of climbing and jumping mentions motivates recordings with more environmental opportunities for these behaviors. Source-time deduplication prevents overlapping event-guided windows from increasing these counts.

Table 13: Behavior-related word mentions in E after source-time deduplication.

Word family	**Mouse slots**	**Mouse (%)**	**Rat slots**	**Rat (%)**
Locomotion	2,339	12.41	269	2.30
Turning	4,371	23.20	1,928	16.51
Upright posture	3,432	18.22	1,428	12.23
Grooming	1,400	7.43	640	5.48
Sniffing	2,066	10.97	179	1.53
Stillness	1,991	10.57	1,895	16.23
Eating or drinking	177	0.94	27	0.23
Climbing or jumping	3	0.02	12	0.10

B.5 Motion Preprocessing

Screening checks observation coverage, bad-frame prevalence, landmark validity, root trajectories, and coordinate transforms. Internal gaps of at most 0.2 seconds are interpolated between reliable endpoints within a continuous region. Each window receives a fixed transform into meters, with the initial horizontal SpineM position at the origin, an approximate ground height from the lowest valid landmark, and the initial heading aligned around the vertical axis. This retains root displacement within the clip and the physical scale difference between mice and rats. Anti-aliasing and resampling produce 30 Hz motion with 60, 75, 90, 105, or 120 frames, preserving the original duration. Windows are padded to $120 \times 25 \times 3$; masks restrict losses and evaluation to their actual lengths and reliable coordinates. Input-validity masks admit accepted interpolations, while supervision masks exclude unreliable landmarks and padding. For joint mouse–rat training, shared coordinate statistics are computed from the pooled physical-coordinate training data.

B.6 Kinematic and Window Details

Relative-motion conventions. In Equation 2, $\mathbf{R}_r$ specifies reference axes, $\mathbf{p}_b$ is the landmark being tracked (e.g., a limb endpoint), $\mathbf{a}_r$ is its reference anchor, and L is the median SpineF–SpineL distance used as the reference body length. Ground-referenced and body-relative measurements share this form: angles are derived from the direction components, while positions describe movement relative to the selected anchor, such as a limb's shoulder or hip. For an angular or positional feature f, the difference between short endpoint intervals summarizes its net change; angles are unwrapped within continuous valid segments before subtraction. Within-interval variation additionally identifies movements that return toward their initial state.

Event-window selection. The algorithm selects at most one window per two-second anchor, using VLM indices to favor coherent actions and informative phase boundaries while accounting for uncertainty. A window can cover part of an ongoing activity.

Candidates with matching semantic signatures are deduplicated at temporal intersection-over-union ≥ 0.75; exact fixed-window duplicates are removed. Windows stay within continuous recordings and reuse source-aligned E and K with timestamps relative to the new window.

Table 14: Source recordings, construction windows, and retained benchmark statistics.

Quantity	Mouse	Rat	Total
Source recordings	5	7	12
Source timepoints	1,116,000	1,161,592	2,277,592
Reconstructed duration (hours)	3.10	2.69	5.79
Construction archive (windows)	7,787	4,933	12,720
Training windows	4,774	2,309	7,083
Validation windows	292	162	454
Test windows	366	292	658
Frozen benchmark windows	5,432	2,763	8,195
B caption entries	16,296	8,289	24,585
E slots	30,881	16,261	47,142
Nonempty E descriptions	30,709	15,970	46,679
Motionless slots	172	291	463
B + nonempty E entries	47,005	24,259	71,264

C Reconstruction Error, Coverage, and Trajectory Diagnostics

Manual annotation sets. We align Rodent25 predictions to 500 annotated frames each from Mouse M1 and Rat s1d1 by recording and sample ID. All 1,000 frames match, yielding 12,500 finite landmark comparisons per species. MPJPE and percentage of correct keypoints (PCK) cover all 25 landmarks in original millimeter coordinates, with no rigid alignment, scale fitting, interpolation, or predicted reference labels. PCK counts the percentage of valid landmark–frame pairs whose position error is at most δ millimeters (5 or 10 in Table 15). For the valid pairs $\mathcal{V}$, the metrics are

$$\mathrm{MPJPE} = \frac{1}{|\mathcal{V}|} \sum_{(t,j)\in\mathcal{V}} \|\widehat{\mathbf{p}}_{tj} - \mathbf{p}_{tj}\|_2, \qquad \mathrm{PCK}_\delta = \frac{100}{|\mathcal{V}|} \sum_{(t,j)\in\mathcal{V}} \mathbb{1}\{\|\widehat{\mathbf{p}}_{tj} - \mathbf{p}_{tj}\|_2 \le \delta\}. \tag{10}$$

Table 15 separates head, trunk, tail, and limb errors. The head contains three landmarks, the trunk three, the tail three, and each paired limb group eight. These anatomical groups partition the 25 landmarks, so the overall error is their landmark-count-weighted mean. The annotation sets include model-development frames, so the errors measure agreement with available labels rather than held-out generalization. Source mocap and Rodent25 have different landmark definitions; their coverage and consistency are compared separately below.

Full-recording audit. The full-recording comparison includes all available source coordinates and 2,277,592 Rodent25 predictions before window screening. The source mouse files contain 310 sparse labeled timepoints across five recordings; their 3D labels are repeated in the camera records and are counted once. Rat-7M provides 1,161,592 mocap timepoints across seven recordings. The source skeletons contain 22 mouse landmarks and 20 rat markers; Rodent25 contains 25 anatomical landmarks, including a three-landmark tail and four landmarks along each limb. The rat source markers include auxiliary offsets and do not define the same anatomical landmarks as Rodent25. Table 16 compares body-part coverage.

Completeness and consistency. For source labels and mocap, a valid landmark observation has three finite coordinates; confidence-screened Rodent25 observations must also pass the frame-level QC rule below. Coverage is the number of such observations divided by the available timepoint–landmark pairs, with no interpolation. For each native limb edge, we compute the within-recording coefficient of variation $100\,\sigma(\ell)/\mu(\ell)$ of its finite lengths, then average across edges and recordings. CV excludes head, torso, and tail spans and uses each resource's native limb edges. It measures consistency within those skeletal definitions, rather than matched-landmark accuracy. Bone geometry and temporal constraints are established tools for examining animal pose trajectories (Karashchuk et al., 2021; Li et al., 2023; Monsees et al., 2022).

Table 15: Overall and body-part reconstruction error on 500 annotated frames each from Mouse M1 and Rat s1d1.

Species	Part	Landmark–frame pairs	MPJPE (mm)	PCK@5 mm (%)	PCK@10 mm (%)
Mouse	All	12,500	2.275	95.66	99.77
	Head	1,500	1.427	99.73	99.73
	Trunk	1,500	2.451	94.80	99.93
	Tail	1,500	1.546	99.00	100.00
	Forelimbs	4,000	2.447	95.80	99.75
	Hindlimbs	4,000	2.628	93.08	99.65
Rat	All	12,500	8.449	36.71	75.68
	Head	1,500	4.590	69.67	94.07
	Trunk	1,500	6.154	43.07	87.60
	Tail	1,500	16.009	30.73	52.47
	Forelimbs	4,000	8.284	29.85	73.98
	Hindlimbs	4,000	8.087	31.08	74.72

Table 16: Landmark coverage by body part; starred limb counts are per side.

Representation	Head	Spine	Forelimb*	Hindlimb*	Tail	Auxiliary
Mouse source (22)	3	2	4	3	3	0
Rat mocap (20)	3	3	3	3	0	2
Rodent25 (25)	3	3	4	4	3	0

Consecutive-coordinate availability. Adj. is the fraction of landmark–frame transitions for which both endpoints are valid and the source timepoints are consecutive. Rodent25 uses the confidence-screened mask in the per-recording tables. The denominator includes all adjacent array entries for every native landmark within each recording; transitions across source-time gaps do not contribute to the numerator. We join mouse sample IDs to the original camera frame map and use saved rat timestamps to identify consecutive observations. Sparse mouse annotations have no Adj. entry. The aggregate Valid and Adj. values in Table 2 pool their respective numerators and denominators across recordings, without joining recording boundaries. CV first averages native limb-edge coefficients within each recording, then averages recordings equally; the main table reports each species separately.

Confidence and screening. For all twelve RodentBEK recordings, the production-QC rule supplements the finite-coordinate check: let s_t be the median of the 25 landmark heatmap responses at frame t, and retain the frame when $s_t > Q_{0.01}(s)$ within its recording. Heatmap responses are uncalibrated, so thresholds are computed per recording. We mask all landmarks in excluded frames, retain the original frame count as the denominator, and count an adjacent pair only when both frames pass and are consecutive in source time. CV is recomputed on retained limb observations with the same recording-level aggregation. The 99% target is an experimental retention setting, not a localization-accuracy estimate.

On the 500-frame annotation sets, the screen excludes four Mouse M1 frames and five Rat s1d1 frames. Retained/rejected MPJPE is 2.267/3.216 mm for mouse and 8.389/14.385 mm for rat, compared with unfiltered means of 2.275 and 8.449 mm. The per-recording thresholds and masks are retained with the audit artifacts.

This coordinate audit applies an additional confidence screen. The generation experiments use the masks in Appendix B.5; mouse masks depend on finite coordinates, bone-length flags, and interpolation provenance.

Recording-level comparison. With equal recording weights, screened limb-length CV is 15.07% for mouse and 29.43% for rat, compared with 14.41% for mouse source labels and 28.54% for rat mocap. The mouse estimates approach the sparse-label reference while covering continuous motion. Four of the seven rat recordings have lower CV than the corresponding mocap records; higher CV in s5d1 and s5d2 raises the equally weighted rat mean above the mocap mean.

Camera coverage and calibration may contribute to the recording-level differences. Mouse M5 uses the available views because Camera1 coverage is incomplete. Rat s5d1 and s5d2 have suspected camera-parameter issues that may affect reconstruction.

Table 17: Mouse coordinate quality: 99% QC retention for Rodent25; finite validity for source labels (†: incomplete Camera1 coverage).

Record	Coordinates	Frames	Valid (%)	CV (%)	Adj. (%)
m1	Labels	81	96.24	13.43	–
m1	Rodent25	18,000	99.00	14.45	98.88
m2	Labels	91	98.25	13.76	–
m2	Rodent25	18,000	99.00	14.72	98.91
m3	Labels	48	98.20	11.18	–
m3	Rodent25	360,000	99.00	12.36	98.81
m4	Labels	44	94.32	18.71	–
m4	Rodent25	360,000	99.00	18.43	98.85
m5	Labels	46	93.87	14.96	–
m5†	Rodent25	360,000	99.00	15.37	98.80

Table 18: Rat coordinate quality: 99% QC retention for Rodent25; finite validity for mocap (‡: suspected calibration issues).

Record	Coordinates	Frames	Valid (%)	CV (%)	Adj. (%)
s1d1	Mocap	126,000	90.52	26.92	89.92
s1d1	Rodent25	126,000	99.00	19.98	98.68
s2d1	Mocap	184,831	93.08	31.66	92.63
s2d1	Rodent25	184,831	99.00	24.88	98.78
s2d2	Mocap	165,375	92.02	49.24	91.27
s2d2	Rodent25	165,375	99.00	24.84	98.73
s3d1	Mocap	177,935	94.05	22.99	93.64
s3d1	Rodent25	177,935	99.00	23.52	98.71
s4d1	Mocap	59,452	91.67	24.28	91.39
s4d1	Rodent25	59,452	99.00	22.67	98.71
s5d1	Mocap	234,499	95.17	25.18	95.04
s5d1‡	Rodent25	234,499	99.00	53.28	98.82
s5d2	Mocap	213,500	95.85	19.55	95.76
s5d2‡	Rodent25	213,500	99.00	36.81	98.79