CAN MOTION-LANGUAGE MODELS GROUND STRUCTURE? STRIDE FOR EVALUATING THE EVALUATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Motion-language models are typically scored by motion-language evaluators, but how well these evaluators ground language structure remains unclear. Here, we introduce the **S**tructure grounding via **T**emporal-order, **R**eflection, and **I**dentity **D**iagnostic **E**valuation (STRIDE) benchmark to systematically evaluate the ability of evaluators to track temporal order, mirror reflection, and action identity. STRIDE comprises 5,869 triples, each consisting of a motion, its original caption, and a perturbed caption, spanning both short and long descriptions. We likelihood-balance caption pairs to reduce text-only bias and estimate each evaluator's caption preference under unrelated motions to measure the discrimination gain from matched motions relative to this baseline. Our experiments reveal weak structural grounding and severe deficits in mirror sensitivity among the audited evaluators, which commonly used evaluation protocols fail to expose. To understand why these limitations go undetected in standard tests, we examine the evaluators more closely. We find that text-only priors alone can solve naive perturbation tests on existing datasets, while common retrieval and distributional metrics barely respond to structural corruption introduced by mirroring ground-truth motions. These findings suggest a natural intervention: structural hard negatives. Our experiments show that a simple modification to contrastive learning substantially improves performance on temporal order and mirror reflection. The benchmark and code will be released.

1 INTRODUCTION

Motion generation and understanding have important applications in embodied intelligence, computer animation, and world-model simulation (Tevet et al., 2023; Zhang et al., 2024; Chen et al., 2023; Zhang et al., 2023b;a; Guo et al., 2022b; 2024). Current evaluation commonly uses pretrained motion-language evaluators to extract embeddings for computing metrics such as R-Precision, Fréchet Inception Distance (FID) (Heusel et al., 2017), and Motion Matching Distance (MM-Dist). A widely used example is the contrastive text-motion embedding evaluator introduced with HumanML3D (Guo et al., 2022a). But does performance on these metrics really indicate that a model understands the correspondence between motion and text? The concern is that, if an evaluator itself cannot distinguish structurally different motions or captions, metrics computed in its embedding space will also struggle to expose the corresponding errors. Figure 1 presents three concrete examples: we fix each real motion, change only temporal order, mirror reflection, or action identity in its caption, and score the resulting pairs with the HumanML3D evaluator. Surprisingly, the evaluator prefers the perturbed, incorrect caption in all three examples. These observations expose failures of the HumanML3D evaluator to distinguish fine-grained structural differences and raise a further question: do other evaluators exhibit similar shortcomings?

Answering this question first requires examining whether standard training and evaluation sufficiently demand structural discrimination. Many motion-language matchers use contrastive training to distinguish paired from unpaired samples, while retrieval evaluates whether they recover the correct pair (Petrovich et al., 2023; Chen et al., 2024b; Guo et al., 2025). For example, given a motion of someone raising their left hand, an evaluator choosing among "raising the left hand," "walking," and "sitting down" may succeed by recognizing hand raising without distinguishing left from right. In contrast, reliably distinguishing "raising the left hand" from "raising the right hand" based on the observed motion requires grounding the left–right distinction in that motion. Thus, when non-

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.

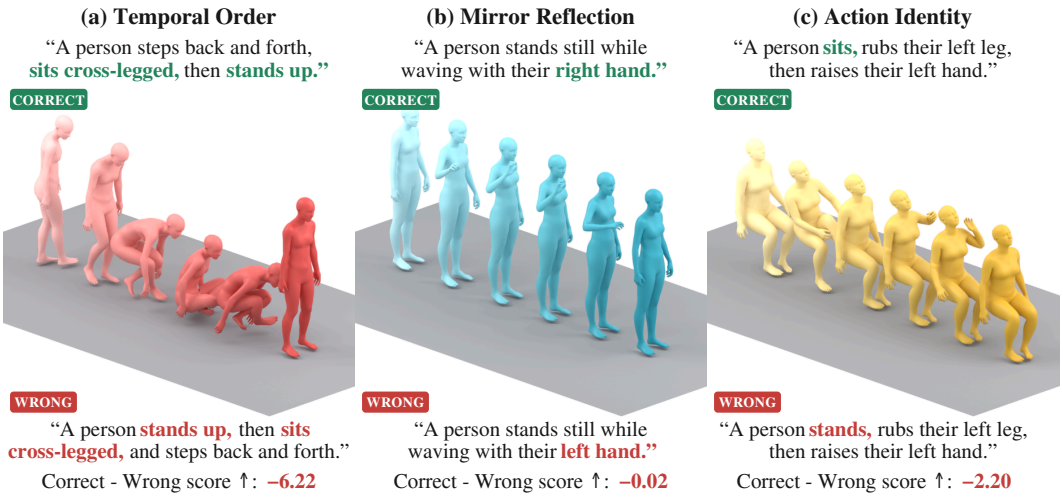


Figure 1: Structural grounding failures of the HumanML3D evaluator. Each panel holds a fixed real motion and perturbs its caption along one axis: temporal order, mirror reflection, or action identity. In all three examples, the correct−wrong score is negative, indicating the evaluator ranks the incorrect caption above the correct one.

matching candidates in training and evaluation differ mainly in overall action content, successful matching need not imply fine-grained structural discrimination (Yuksekgonul et al., 2023). However, even when candidate captions differ only in the target structure, a correct choice does not necessarily indicate that the evaluator used motion information: it may still rely on fluency or caption preference (Hsieh et al., 2023). We therefore introduce STRIDE, a diagnostic benchmark covering temporal order, mirror reflection, and action identity, which uses caption preference under unrelated motions as a baseline to assess whether matched motions improve evaluators' discrimination of structural differences. Table 1 compares STRIDE with existing evaluation protocols in structural coverage and preference control.

Table 1: Comparison with existing motion-language evaluation protocols. STRIDE combines controlled tests of temporal order, mirror reflection, and action identity with text-only bias checks and evaluator-specific caption-preference controls across both short and long descriptions.

Evaluation Setting	Content Matching	Temporal Order	Mirror Reflection	Action Identity	Matched Pairs	Prior Controls	Short Captions	Long Captions
HumanML3D suite (Guo et al., 2022a)	✓	–	–	–	–	–	✓	–
TMR protocols (Petrovich et al., 2023)	✓	–	–	–	–	–	✓	–
CLaM (Chen et al., 2024b)	✓	–	–	–	–	–	✓	–
MoBERT (Voas et al., 2023)	–	–	–	–	–	–	✓	–
ChronRet (CAR) (Fujiwara et al., 2024)	✓	✓	–	–	✓	–	✓	–
SnapMoGen suite (Guo et al., 2025)	✓	–	–	–	–	–	–	✓
STRIDE (ours)	✓	✓	✓	✓	✓	✓	✓	✓

Our contributions are threefold:

- We introduce STRIDE, a benchmark for fine-grained evaluation of structural grounding in motion-language evaluators, with controlled tasks covering temporal order, mirror reflection, and action identity across short and long captions. We combine caption-preference controls with a Random Content reference to assess capabilities along each structural axis.
- We systematically evaluate six released motion-language evaluators and reveal a gap between overall content matching and fine-grained structural discrimination. After controlling for caption preference under unrelated motions, all tested evaluators achieve lower grounding scores on each structural axis than on their Random Content reference, with Mirror Reflection a recurring weakness for most evaluators.

- We analyze why contrastive training may not require fine-grained structural discrimination when negatives differ mainly in overall content. Based on this analysis, we introduce a structure-aware hard-negative scheme. Experiments show that this scheme can improve temporal-order and mirror-reflection grounding in the tested evaluators, with gains varying across evaluators and structural axes.

2 THE STRIDE BENCHMARK: WHEN DO EVALUATORS MISS STRUCTURE?

As shown in Figure 2, we introduce STRIDE, a benchmark built from HumanML3D and SnapMoGen to validate whether motion-language evaluators recognize structural differences. We first define the Temporal Order, Mirror Reflection, and Action Identity perturbations (Section 2.1), then describe benchmark construction and quality checks (Section 2.2). Next, we introduce unrelated-motion controls to measure the discrimination gain from matched motions over the caption-preference baseline (Section 2.3). Finally, we compare released evaluators across structural axes and datasets (Section 2.4).

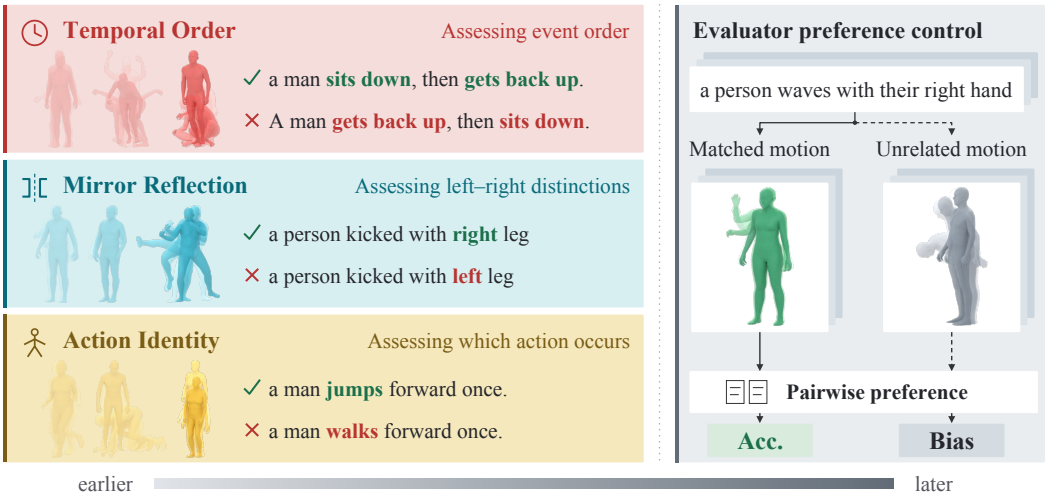


Figure 2: STRIDE probes three structural distinctions and measures each evaluator's caption preference under an unrelated-motion control. Left: each dark motion sequence is paired with its original and edited captions; highlighted phrases mark the structural change, and lighter-to-darker shading encodes relative time. Right: green motion sequences match the caption, whereas grey sequences are unrelated to it and serve to estimate model preference. Each evaluator assesses motion-caption matching under the matched- and unrelated-motion conditions to obtain acc and bias, respectively.

2.1 STRUCTURAL PERTURBATIONS

We construct structural perturbations by changing one target property of a motion caption while keeping the motion and the other caption information fixed. STRIDE tests three types of changes: **Temporal Order** changes the order of events in a multi-action caption; **Mirror Reflection** exchanges left and right in the description; and **Action Identity** replaces one action with a visually distinct alternative. These tests ask whether evaluators distinguish specific structural differences. To compare structural discrimination with overall content matching, we replace the matched caption with the caption of a different corpus motion as a reference of **Random Content** matching. Each STRIDE item is a triplet $i = (m, c, c')$: a real motion m, its ground-truth caption c, and a single-property perturbation c'. Given m, the evaluator chooses between c and c'.

Temporal edits preserve the original actions, actor, and scene. Mirror edits replace every directional description (e.g., *left→right*), preserving fluency and corresponding to a physically valid mirrored motion. Action Identity edits use predefined verb pairs (e.g., *runs→walks*) while preserving grammatical form and arguments, such as the subject and objects.

2.2 BENCHMARK CONSTRUCTION

The benchmark uses the official HumanML3D and SnapMoGen test splits, which provide short clip-level captions and long-form descriptions, respectively.

We construct perturbations from each ground-truth caption. For Temporal Order, we reverse the event order in multi-action descriptions while retaining every action. For Mirror Reflection, we swap all occurrences of *left* and *right*. For Action Identity, we replace one verb using the predefined action pairs while preserving its grammatical form. Action Identity candidates also undergo an independent text-only check, with no motion shown. Implementation details are provided in Appendix A.

An edited caption may be less fluent than the original, allowing an evaluator to choose correctly from text fluency. For example, a Temporal Order perturbation changes "a person is waving their right arm up and down and then side to side" to "a person side to side then is waving their right arm up and down." The missing verb in the opening clause provides a grammatical cue without requiring the evaluator to inspect the motion. We limit this difference with a likelihood filter. With $\bar{\ell}_L(c)$ the mean token log-likelihood of caption c under GPT-2 Large (Radford et al., 2019), the caption-prior gap is defined as follows:

$$\Delta\ell\ell(c, c') = \bar{\ell}_L(c) - \bar{\ell}_L(c') \tag{1}$$

where positive values favor the original caption from text alone; structural pairs are kept when $|\Delta\ell\ell| \leq 0.15$ nats per token.

To check whether the constructed caption pairs still contain exploitable text-only cues, we use GPT-2, Vera, and Grammar to examine gaps in language likelihood, commonsense plausibility, and grammatical acceptability, respectively, and compare STRIDE with rule-based perturbations. As shown in Figure 3, STRIDE's score gaps on Temporal Order and Action Identity are more concentrated near zero, indicating reduced text-only preferences in some aspects of the constructed data.

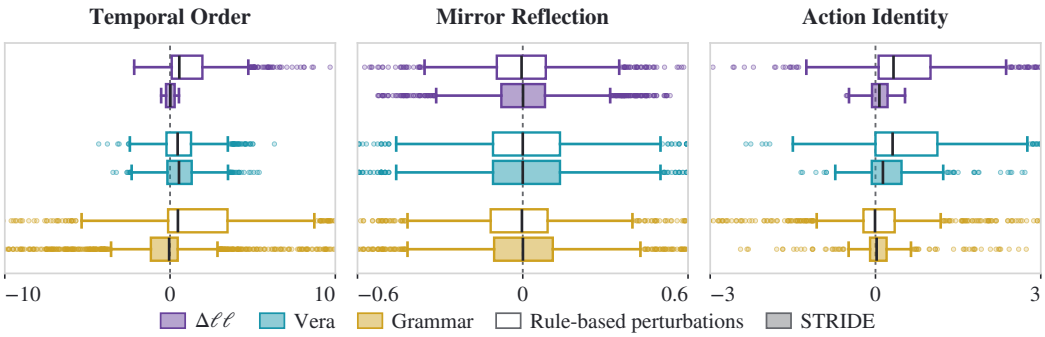


Figure 3: Comparison of text-only score gaps between rule-based perturbations and STRIDE. Colors distinguish $\Delta\ell\ell$, Vera, and Grammar; open and filled boxes represent rule-based perturbations and STRIDE, respectively. Gaps are original minus perturbed scores, scaled by each scorer's interquartile range (IQR) on the rule-based structural pairs; zero indicates equal scores. The two pools are not matched before/after samples.

Finally, a human audit assesses item quality in STRIDE by checking caption fluency, whether only one property has changed, and whether the change can be distinguished from the motion. The released benchmark contains 4,269 structural pairs plus 1,600 Random Content references: 5,869 triples over 2,598 motions.

Figure 4 shows word frequencies by axis and motion-coverage statistics for 5,000 clips sampled from each source corpus. The corpora overlap broadly in the shared kinematic projection, with SnapMoGen more evenly spread across the grid. Appendix A.2 details the analysis.

2.3 EVALUATION WITH UNRELATED-MOTION CONTROLS

Controlled caption pairs focus the test on the targeted distinction, but a correct choice alone does not establish that the matched motion helped. We therefore ask a further question: does the evaluator distinguish the same captions more reliably with the matched motion than with an unrelated motion?

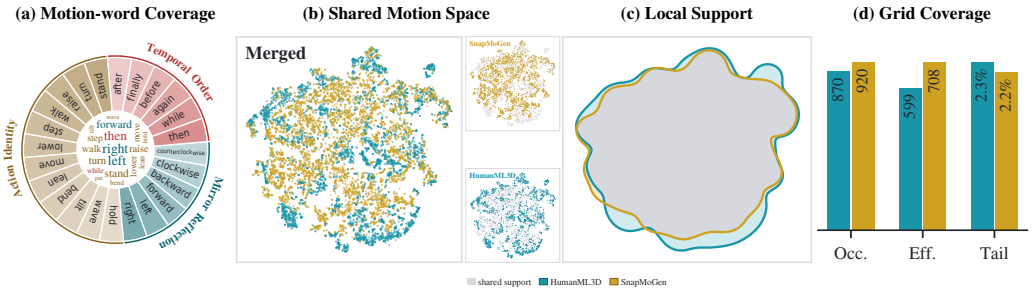


Figure 4: Vocabulary and motion coverage. (a) Frequent words by axis. (b) Shared-feature t-SNE projection. (c) Smoothed corpus footprints. (d) Occupied cells (Occ.), effective cells (Eff.), and long-tail share (Tail) on a 55×55 grid; bars are normalized per metric, with unscaled labels.

Raw accuracy can reflect both motion discrimination and caption preference: if original captions are more natural, an evaluator can score above chance without using the motion. To measure the actual gain from the observed motion, we keep the captions fixed and evaluate them with matched and unrelated motions. For item i, let d_i and b_i denote the original-caption score minus the perturbed-caption score under the matched motion and the unrelated motion $\tilde{m}_i$, respectively. Each condition is scored separately: positive, zero, and negative margins receive 1, 0.5, and 0, respectively. For each evaluator, we average these values separately within each corpus and axis to obtain matched-motion accuracy acc and the original-caption preference rate under unrelated motions, bias. We measure the accuracy gain from matched motions as $\mathrm{acc} - \mathrm{bias}$ and divide it by the remaining room for improvement, $1 - \mathrm{bias}$, to obtain the normalized gain:

$$\Delta = \frac{\mathrm{acc} - \mathrm{bias}}{1 - \mathrm{bias}} \tag{2}$$

where Δ denotes the ability to recognize the structural gap between matched motions and captions. Values of Δ closer to 1 indicate better performance.

While Δ measures the overall accuracy gain, an improvement in scores need not change the final choice. To capture such improvements, we propose *Grounded Perturbation Recognition* (GPR). GPR measures the proportion of items whose original-minus-perturbed caption margin is larger under the matched motion than under the unrelated motion. GPR is calculated as follows:

$$\mathrm{GPR} = \Pr[d_i > b_i] + \tfrac{1}{2}\Pr[d_i = b_i] \tag{3}$$

where a GPR closer to 100% indicates that matched motions push scores toward the correct structural interpretation on more items.

To examine the need for these controls, Figure 5 reports each model's original-caption preference rate (bias) under unrelated motions. Even after data-side filtering, these rates vary across models and perturbation axes, indicating that the evaluated models still exhibit some caption preferences under this control.

	HumanML3D eval.	TMR	CLaM	MoBERT	ChronRet	SnapMoGen eval.	MotionGPT
Temporal Order	39.2	49.4	44.1	49.5	50.3	51.8	89.4
Mirror Reflection	48.4	52.7	51.9	50.1	51.5	51.9	51.4
Action Identity	50.0	55.7	53.8	53.3	58.5	52.9	67.9
Random Content	50.3	50.6	49.0	47.6	51.4	50.0	48.8

Figure 5: Comparison of caption preferences across models under unrelated motions. Cells report the rate of preferring the original over the perturbed caption (bias, %; ties count as half). White marks 50%; purple and teal indicate preferences for the original and perturbed captions, respectively.

2.4 Evaluation Results

Using these unrelated-motion preferences as baselines, we next examine whether matched-motion gains in structural discrimination remain below those in overall content matching, and which eval-

uators and axes show the largest gaps. Our primary audit concerns released evaluators; adapted generative and vision-language interfaces are examined separately as supplementary tests.

Table 2 audits six released evaluators: the HumanML3D block holds the HumanML3D evaluator, TMR, CLaM, MoBERT, and the chronology-trained ChronRet on their native representation, and the SnapMoGen block holds the native SnapMoGen evaluator plus the five HumanML3D-trained evaluators. Table 2 reports acc, bias, and Δ per axis, plus macro GPR.

Table 2: Performance comparison of existing motion-language evaluators on STRIDE. We report matched-motion accuracy (acc), original-caption preference under unrelated motions (bias), and normalized grounding gain Δ; acc and bias are reported as proportions. Macro GPR (%) averages GPR over the three structural axes, while Random Content serves as a content-matching reference. Bold and underlined values denote the best and second-best Δ and GPR within each dataset block, respectively. Appendix C details model inputs and evaluation settings.

	Temporal Order			Mirror Reflection			Action Identity			Random Content			Macro
Published evaluator	acc	bias	$\Delta\uparrow$	acc	bias	$\Delta\uparrow$	acc	bias	$\Delta\uparrow$	acc	bias	$\Delta\uparrow$	GPR (%)↑
HumanML3D (native input)													
HumanML3D evaluator (Guo et al., 2022a)	0.624	0.392	0.381	0.463	0.484	-0.041	0.858	0.500	0.717	0.940	0.503	0.879	66.4
TMR (Petrovich et al., 2023)	0.704	0.494	0.416	0.728	0.527	0.424	0.887	0.557	0.745	0.955	0.506	0.910	70.7
CLaM (Chen et al., 2024b)	0.726	0.441	**0.510**	0.737	0.519	**0.454**	0.925	0.538	**0.837**	0.978	0.490	**0.956**	**73.9**
MoBERT (Voas et al., 2023)	0.726	0.495	0.457	0.594	0.501	0.186	0.844	0.533	0.667	0.937	0.476	0.880	70.3
ChronRet (Fujiwara et al., 2024)	0.698	0.503	0.392	0.693	0.515	0.367	0.925	0.585	0.818	0.970	0.514	0.938	72.4
SnapMoGen (long captions)													
HumanML3D evaluator (Guo et al., 2022a)	0.541	0.503	0.076	0.513	0.504	0.018	0.564	0.505	0.119	0.748	0.516	0.478	52.5
TMR (Petrovich et al., 2023)	0.510	0.466	0.082	0.617	0.529	0.187	0.642	0.456	0.342	0.789	0.520	0.560	57.7
CLaM (Chen et al., 2024b)	0.569	0.478	0.174	0.600	0.488	0.219	0.642	0.510	0.270	0.848	0.535	0.672	56.9
MoBERT (Voas et al., 2023)	0.552	0.501	0.102	0.534	0.499	0.070	0.547	0.493	0.106	0.733	0.535	0.425	55.8
ChronRet (Fujiwara et al., 2024)	0.568	0.496	0.143	0.599	0.543	0.123	0.593	0.549	0.098	0.816	0.502	0.631	58.8
SnapMoGen evaluator (Guo et al., 2025)	0.965	0.518	**0.927**	0.728	0.519	**0.435**	0.838	0.529	**0.656**	0.998	0.500	**0.995**	**74.7**

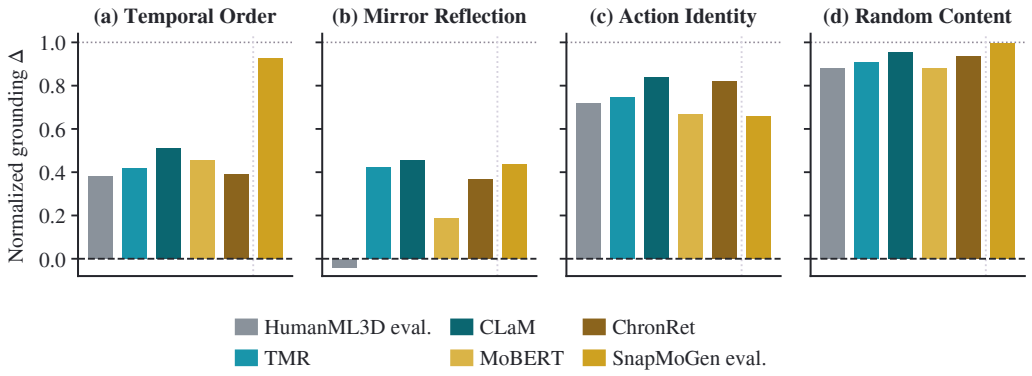


Figure 6: Performance comparison of existing evaluators across perturbation axes. Bars report normalized grounding Δ (higher is better) from Table 2 for the three structural axes and the Random Content reference. In each panel, the first five evaluators are tested on HumanML3D; the SnapMoGen evaluator, to the right of the vertical separator, is tested on SnapMoGen. Comparisons across the separator are not same-dataset rankings. The horizontal dashed and dotted lines mark $\Delta = 0$ (no gain over unrelated motion) and $\Delta = 1$ (perfect matched-motion accuracy), respectively.

Table 2 shows that every evaluator obtains lower Δ scores on all three structural axes than on its Random Content reference, revealing a clear gap between content matching and structural discrimination. Figure 6 visualizes this gap for each evaluator on its native dataset. For most evaluators, Mirror Reflection has the lowest Δ of the three structural axes, indicating a recurring weakness in left–right discrimination. On HumanML3D, Temporal Order and Mirror Reflection are particularly weak: even the best-performing evaluator reaches Δ scores of only 0.510 and 0.454, respectively. Action Identity is comparatively easier to distinguish. On SnapMoGen, commonly used evaluators trained on short captions perform still worse with the dataset's long captions. The SnapMoGen evaluator, however, shows a clear advantage on Temporal Order, achieving a Δ of 0.927. This

may benefit from the motion reconstruction constraint used during evaluator training. In addition to cross-modal alignment, this constraint requires both text and motion embeddings to reconstruct the corresponding motion sequence, which may encourage the representations to preserve action order and help detect temporal mismatches. Nevertheless, its Mirror Reflection and Action Identity Δ scores remain well below its Random Content reference.

Macro GPR complements the Δ analysis at the sample level by measuring whether matched motions consistently increase the original caption's score advantage over its perturbed counterpart. All tested evaluators have a macro GPR below 75%, with even the highest-scoring SnapMoGen evaluator reaching only 74.7%. This indicates that, for a substantial fraction of structurally perturbed caption pairs, the tested evaluators fail to use matched motions to increase the original caption's score advantage over its perturbed counterpart.

Evaluations on adapted generative and vision-language interfaces also show that high caption-selection accuracy can coexist with strong caption preferences under unrelated motions, while the keyframe-based VLM results are limited by weak overall content matching, details are in Appendix D.2.

3 Why can structural errors be missed?

STRIDE exposes weaknesses in structural discrimination, but why are these weaknesses not fully reflected in standard evaluation metrics? To examine this question, we investigate whether these limitations affect the evaluators' use in motion-generation evaluation. We shift the intervention to the motion side and first test whether standard metrics detect structural errors in the motions. We then discuss why the matching objective in contrastive training may provide insufficient incentive to learn these distinctions.

3.1 Limitations of retrieval-based evaluations

To isolate the metrics' response to structural errors, we start from ground-truth motions, keep their captions fixed, and apply controlled motion transformations before scoring the resulting pairs with the standard generation-evaluation suite. This avoids confounding the analysis with other quality differences between outputs from different generators. Table 3 compares the original ground-truth motions, mirrored motions, time-reversed motions, and a control with shuffled motion-caption pairings. Time reversal also changes local dynamics, so it is not equivalent to STRIDE's caption perturbations that change only event order.

Table 3: Comparison of standard motion-generation metrics under motion-side interventions. We compare mirroring, time reversal, and shuffled pairing against the original ground-truth reference. R@K denotes R-Precision at K, reported as a proportion; FID and MM-Dist denote Fréchet Inception Distance and motion-text matching distance, respectively.

Motion Condition	R@1↑	R@2↑	R@3↑	FID↓	MM-Dist↓	Diversity→
Original Motion	0.529	0.710	0.812	0.000	2.914	9.272
Mirror Reflection	0.523	0.713	0.815	0.020	2.904	9.777
Time Reversal	0.357	0.523	0.632	8.906	4.729	8.327
Shuffled Pairing	0.032	0.066	0.098	0.000	9.516	9.349

Table 3 shows that, although the matching metrics respond clearly to Time Reversal and Shuffled Pairing, mirroring changes R@1 only from 0.529 to 0.523, and even yields a slightly better MM-Dist score than the original motions. These results indicate that standard evaluation scores may fail to reveal left–right structural mismatches between motions and captions. Under the evaluated protocol and dataset, good overall matching scores can thus be obtained without correctly distinguishing left from right, allowing these errors to remain hidden in the final scores.

3.2 LIMITATIONS OF CONTRASTIVE LEARNING-BASED TRAINING

The preceding results explain how structural weaknesses can go unnoticed in evaluation, but not why they arise. For contrastively trained motion-language matchers, one possible explanation lies in the combination of the objective and the available negatives. The objective requires matched motions and captions to outrank alternatives. If those alternatives differ mainly in overall action content, coarse content matching may suffice to reduce the loss without resolving left–right distinctions or event order.

The issue is therefore not simply whether captions contain structural information, but whether training requires that information to distinguish a correct pair from its negatives. When alternatives with similar content but different structure are missing, ignoring structure may still satisfy the matching objective. This is a possible explanation for contrastive matchers, not an isolated causal finding or a claim about every tested model's training objective.

4 A SIMPLE FIX: STRUCTURAL HARD NEGATIVES

The preceding analysis suggests that content-level negatives may not require fine-grained structural distinctions. We therefore add content-similar, structurally different hard negatives to test whether they improve structural grounding. This approach requires no changes to the evaluator architecture; we refer to it as Neg Suite.

Figure 7 illustrates our structural hard-negative approach. We retain the ordinary negatives used in standard contrastive learning and introduce perturbed captions and motions. We use an asymmetric design: perturbed captions are compared with the original motion, other motions in the batch, and the corresponding perturbed motion, whereas perturbed motions are compared only with their corresponding original and perturbed captions. This design aims to focus the additional motion-side gradients on distinguishing structural errors, avoiding duplicating training signals for overall content mismatches.

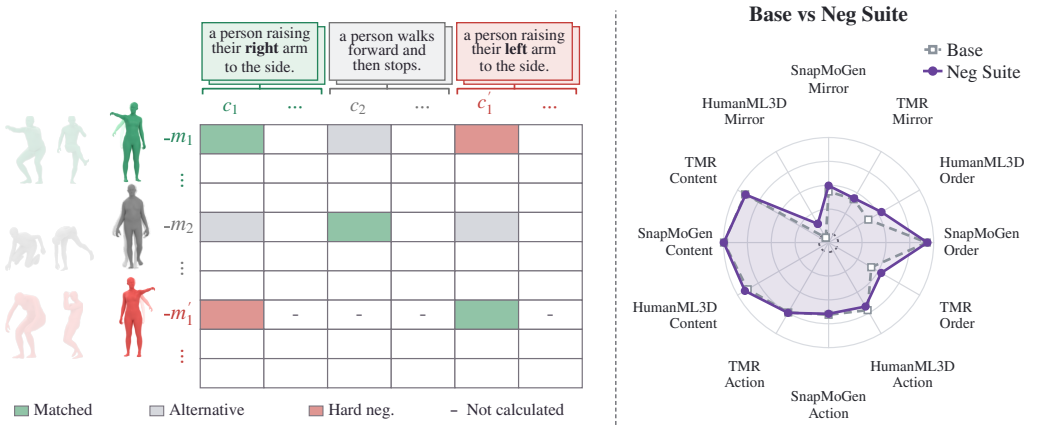


Figure 7: Finetuning with structural hard negatives. Left: correct motion–caption pairs, ordinary negatives, and structural hard negatives. The added motion-side supervision compares only an original pair and its perturbed versions. Right: normalized grounding Δ on the three structural axes and the Random Content control for HumanML3D, TMR, and SnapMoGen, comparing the original evaluators (Base) with their hard-negative-trained variants (Neg Suite).

We test HumanML3D, TMR, and SnapMoGen, comparing normalized grounding Δ and checking retrieval degradation. Appendix E gives negative construction, variant-specific losses, and training settings.

Figure 7 shows per-axis Δ. The gains are most evident in Temporal Order and Mirror Reflection. NegTMR raises Δ on these axes from 0.416 and 0.424 to 0.536 and 0.435, respectively; NegSnap raises them from 0.927 and 0.435 to 0.935 and 0.496. Across TMR and SnapMoGen, five of the six structural scores improve. The HumanML3D evaluator's Mirror Reflection Δ also rises

from -0.041 to 0.127. Structural hard negatives can therefore improve some structural distinctions. Gains on Action Identity are limited, with declines for some evaluators, possibly because these hard negatives resemble ordinary alternative captions and provide less additional supervision.

5 Related work

Motion-language evaluation. HumanML3D established the common evaluation space for text-to-motion generation (Guo et al., 2022a). MotionCLIP aligns motion with the CLIP text space (Tevet et al., 2022); TMR extends TEMOS toward cross-modal retrieval with a follow-up cross-dataset study (Petrovich et al., 2022; 2023; Bensabath et al., 2024); CLaM offers a CLIP-style evaluator (Chen et al., 2024b), SnapMoGen a long-caption T5 evaluator (Guo et al., 2025), and MoBERT a human-calibrated quality regressor (Voas et al., 2023). MotionGPT and MotionLLM model motion-text correspondence through conditional generation (Jiang et al., 2023; Chen et al., 2024a). High clip-level or distributional scores do not guarantee structural discrimination; STRIDE directly tests whether evaluators recognize these structural changes.

Fine-grained and temporal grounding. CAR retrieves the original description against an event-shuffled alternative and trains on such negatives (Fujiwara et al., 2024). FineMoLA learns frame-phrase alignment from clip-level supervision (Wang et al., 2026), and SGAR aligns full-body with part-level motion through body-part concepts (Zhang et al., 2025). Outside motion, video-language diagnostics probe time sensitivity (Bagad et al., 2023; Liu et al., 2024; Li et al., 2024), and self-supervised recognition treats temporal direction and mirror chirality as learnable signals (Misra et al., 2016; Wei et al., 2018; Lin et al., 2020). These works make structure a learning target. STRIDE compares three structural axes against Random Content matching across short and long captions, separately controls data-side text-only bias and unrelated-motion caption preference, and tests hard-negative training without architectural changes.

Compositionality diagnostics and hard-negative learning. Winoground tests composition with curated pairs, and ARO scales relation, attribution, and order diagnostics while showing that targeted negatives improve contrastive models (Thrush et al., 2022; Yuksekgonul et al., 2023); later probes cover compositional retrieval, spatial relations, and object-attribute checklists (Ma et al., 2023; Ray et al., 2023; Kamath et al., 2023; Zhao et al., 2022). SugarCrepe then demonstrates that text-only priors can compromise perturbation benchmarks, and SugarCrepe++ adds sensitivity tests under meaning-preserving lexical variation (Hsieh et al., 2023; Dumpala et al., 2024), echoing language-prior balancing in VQA (Goyal et al., 2017). Building on these compositionality and bias diagnostics, STRIDE asks whether evaluator judgments reflect motion structure, and whether standard generation metrics detect structural mismatches.

6 Conclusion

In this work, we introduce STRIDE, revealing a gap between content matching and structural discrimination in motion-language evaluators. Across short and long captions, all tested released evaluators obtain lower grounding scores on structural axes than on their Random Content reference. Our experiments show that standard generation metrics can miss left–right structural errors. Persistent caption preferences under unrelated motions also show that selecting the correct caption alone does not establish that a model uses motion. One possible explanation for this structural capability gap is that, for contrastive matchers, the training objective may not require fine-grained structural distinctions when negatives differ mainly in overall content. Motivated by this hypothesis, we add structural hard negatives on the caption and motion sides to test targeted supervision, observing improvements in Temporal Order and Mirror Reflection. Future work could further examine the relationship between training objectives and structural grounding, and extend the tests to more models and structures such as motion duration and body-part structure. Motion-language evaluation should complement retrieval and distributional metrics with targeted structural tests. Structural understanding should not be inferred from aggregate scores alone. Such tests can further reveal the strengths and weaknesses of motion-language models and are key to understanding where these models are applicable.

AI USAGE STATEMENT

We used AI tools to assist with writing and polishing, literature retrieval, and generating synthetic caption data. Specifically, AI assisted with Chinese–English translation, language polishing, related-work discovery, and reference checking. In dataset construction, generative AI was used only for caption text: a large language model generated Temporal Order perturbations, and human audits assessed construction quality and text-only preference. The authors take full responsibility for developing the research plan, carrying it out, and the final research decisions, interpretations, references, and manuscript content.

ETHICS STATEMENT

We obtained HumanML3D via AMASS and SnapMoGen under their respective academic licenses. Both datasets contain skeletonized motions without personally identifiable information. Released artifacts contain only captions, perturbed captions, clip identifiers, and code. No motion data is redistributed; reproduction requires agreeing to the original licenses with the respective providers. Human annotation was limited to non-sensitive judgments of motion-caption pairs. Five volunteer annotators aged 22–30 were informed of the study purpose and the intended use of their responses before participation. No identifying or sensitive information was collected. Following consultation under our institutional guidelines, the protocol was determined not to require IRB review.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will publicly release the STRIDE benchmark data and the code for reproducing the experiments and constructing the benchmark. The release will include all $5{,}869$ caption pairs with motion clip identifiers, per-pair bidirectional mirror score grids, training negatives, and construction, evaluation, training, and analysis scripts. The original motion data must be obtained from the dataset providers under their respective licenses and will not be redistributed.

REFERENCES

Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Léore Bensabath, Mathis Petrovich, and Gül Varol. A cross-dataset study for text-based 3D human motion retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. doi: 10.1109/CVPRW63382.2024.00199.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. MotionLLM: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024a.

Xiaodong Chen, Kunlang He, Wu Liu, Xinchen Liu, Zheng-Jun Zha, and Tao Mei. CLaM: An open-source library for performance evaluation of text-driven human motion generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024b. doi: 10.1145/3664647.3685523.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, 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*, 2023.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. SugarCrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Kent Fujiwara, Mikihiro Tanaka, and Qing Yu. Chronologically accurate retrieval for temporal grounding of motion-language models. In *European Conference on Computer Vision*, pp. 323–339, 2024. doi: 10.1007/978-3-031-73636-0_19.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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, 2022a.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts. In *European Conference on Computer Vision*, 2022b.

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.

Chuan Guo, Inwoo Hwang, Jian Wang, and Bing Zhou. SnapMoGen: Human motion generation from expressive texts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 110778–110794, 2025. doi: 10.52202/085713-3341.

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.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.568.

Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. VITATECS: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72897-6_19.

Yuan-Ming Li, Qize Yang, Nan Lei, Shenghao Fu, Ling-An Zeng, Jian-Fang Hu, Xihan Wei, and Wei-Shi Zheng. IRG-MotionLLM: Interleaving motion generation, assessment and refinement for text-to-motion generation. In *European Conference on Computer Vision*, 2026.

Zhiqiu Lin, Jin Sun, Abe Davis, and Noah Snavely. Visual chirality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.517.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, 2016.

Mathis Petrovich, Michael J Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pp. 480–497, 2022.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. Cola: A benchmark for compositional text-to-image retrieval. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. In *European Conference on Computer Vision*, 2022.

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

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Jordan Voas, Yili Wang, Qixing Huang, and Raymond Mooney. What is the best automated metric for text to motion generation? In *ACM SIGGRAPH Asia 2023 Conference Papers*, 2023. doi: 10.1145/3610548.3618185.

Tongyan Wang, Zhengyuan Li, Muhan Lin, Shengyang Luo, Yifan Shen, Aniket Bera, Baijian Yang, and Yingjie Victor Chen. FineMoLA: Towards fine-grained motion-language alignment from clip-level supervision. *arXiv preprint arXiv:2608.01392*, 2026.

Donglai Wei, Joseph Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00840.

Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. MG-MotionLLM: A unified framework for motion comprehension and generation across multiple granularities. *arXiv preprint arXiv:2504.02478*, 2025.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023.

Jiahang Zhang, Lilang Lin, Shuai Yang, and Jiaying Liu. SGAR: Structural generative augmentation for 3D human motion retrieval. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3566.

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, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023b.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024. doi: 10.1109/TPAMI.2024.3355414.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. VL-CheckList: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. *arXiv preprint arXiv:2506.24086*, 2025.

A BENCHMARK CONSTRUCTION AND COVERAGE

A.1 CONSTRUCTION STEPS

Figure 8 summarizes the 5,869 pairs by corpus and axis. We construct structural pairs as follows.

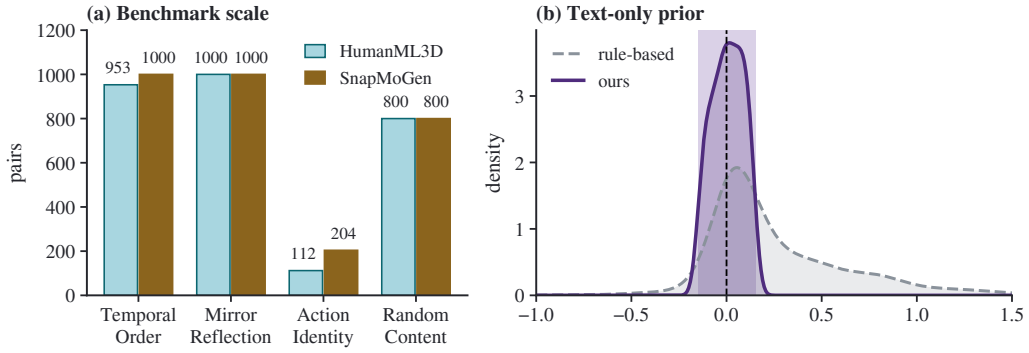


Figure 8: Benchmark composition and likelihood balancing. (a) Pair counts by axis and corpus. (b) Smoothed $\Delta\ell\ell$ densities before and after control. Positive values indicate preference for the original caption; the shaded region marks the acceptance interval.

Generate edits. Temporal Order uses an LLM to rewrite captions with the event order reversed. We use the open-source Qwen3-VL-8B (Bai et al., 2025), instructing it to retain all actions, minimize wording changes, and produce fluent text. Mirror Reflection can be constructed through direct word substitutions, such as swapping left/right or clockwise/counterclockwise. Action Identity replaces one verb using visually distinguishable action pairs, avoiding substitutions such as replacing run with jog, which are difficult to distinguish visually.

Check candidates. Order edits must preserve content words and pass structural checks. Action edits reject candidates in which the source or target action appears more than once. Human reviewers check fluency, whether the actions are distinct and mutually exclusive, and whether only one predicate has changed.

Filter and balance. To reduce the influence of text-only priors, we retain only structural pairs satisfying $|\Delta\ell\ell| \leq 0.15$ nats/token. We then balance Action Identity pairs so that each action pair has equal numbers of examples in both replacement directions. This prevents an action word from consistently appearing in either the correct or the incorrect caption and providing a lexical shortcut.

A.2 VOCABULARY AND MOTION COVERAGE

Figure 4 summarizes vocabulary and motion coverage. We represent 5,000 clips per corpus with 14 body-height-normalized kinematic statistics and visualize both corpora in a shared t-SNE projection. Their distributions overlap broadly, with SnapMoGen more evenly spread. The smoothed coverage regions in Figure 4(c) overlap substantially, with local differences at their boundaries, indicating that the corpora cover similar kinematic regions in this projection. In Figure 4(d), Occ. counts nonempty cells, Eff. measures effective coverage through exponentiated grid-distribution entropy, and Tail measures the sample share in sparse cells containing only one or two samples across both corpora. With equal sample sizes, SnapMoGen occupies slightly more cells (920 vs. 870) and has more effective cells (708 vs. 599), indicating a more even spread within these regions.

B HUMAN AUDITS

We perform two types of human audits. The text-only preference audit examines whether annotators still prefer the original caption without seeing the motion. Five annotators assess text-only caption

pairs, choosing the original caption, the perturbed caption, *equally likely*, or *uncertain*. The audit yields an original-caption preference rate of 56.3% for the resulting benchmark.

The construction quality audit checks whether caption edits meet the intended construction requirements. All judgments must agree that the captions are fluent, only one structural element has changed, and, when motion is shown, the change is visually distinguishable.

C Evaluation implementation

Evaluators use their released checkpoints and native preprocessing.

1. HumanML3D evaluator (Guo et al., 2022a): We use the HumanML3D-trained GloVe–bidirectional-GRU evaluator. The implementation originates from `https://github.com/EricGuo5513/text-to-motion`.
2. TMR (Petrovich et al., 2023): We use the HumanML3D-trained variants. Code and weights are available through `https://github.com/Mathux/TMR`.
3. CLaM (Chen et al., 2024b): We use the released HumanML3D-synthesis evaluator. Code and checkpoint downloads are provided at `https://github.com/SheldongChen/CLaM`.
4. MoBERT (Voas et al., 2023): We use the released `std_bpe2000` primary evaluator. Code and weights are available through `https://github.com/jvoas655/MoBERT`.
5. ChronRet (CAR) (Fujiwara et al., 2024): We use the HumanML3D-trained variants. Code is available at `https://github.com/line/ChronAccRet`, with released weights at `https://huggingface.co/line-corporation/ChronAccRet`.
6. SnapMoGen evaluator (Guo et al., 2025): We use the SnapMoGen-trained six-layer evaluator `eval_klde-5_late-5_nlayer6_norm`, with a T5-v1.1-base text backbone. Code and checkpoint downloads are provided at `https://github.com/snap-research/SnapMoGen`.

For the five HumanML3D-trained evaluators, SnapMoGen's 148-dimensional, 30-fps, 24-joint input is decoded with released kinematics, mapped to 22 joints, resampled to 20 fps, and encoded as 263-dimensional HumanML3D features.

D Additional experiments

D.1 Released-evaluator grounding results

Table 4 complements the main-text grounding results with per-axis GPR.

Table 4: Per-axis GPR (%; chance 50%). Bold and underline mark the best and second-best values within each corpus. † denotes fixed-adapter cross-domain transfer.

Published evaluator	Temporal Order ↑	Mirror Reflection ↑	Action Identity ↑	Macro↑
HumanML3D (native input)				
HumanML3D evaluator (Guo et al., 2022a)	66.9	48.4	<u>84.0</u>	66.4
TMR (Petrovich et al., 2023)	65.5	<u>66.6</u>	80.2	70.7
CLaM (Chen et al., 2024b)	<u>68.1</u>	64.0	**89.6**	**73.9**
MoBERT (Voas et al., 2023)	**69.7**	58.5	82.5	70.3
ChronRet (Fujiwara et al., 2024)	67.4	**66.9**	83.0	<u>72.4</u>
SnapMoGen (long captions)				
HumanML3D evaluator† (Guo et al., 2022a)	53.6	49.4	54.4	52.5
TMR† (Petrovich et al., 2023)	54.5	57.8	<u>60.8</u>	57.7
CLaM† (Chen et al., 2024b)	58.9	57.7	53.9	56.9
MoBERT† (Voas et al., 2023)	55.5	52.3	59.6	55.8
ChronRet† (Fujiwara et al., 2024)	<u>59.0</u>	<u>58.0</u>	59.3	<u>58.8</u>
SnapMoGen evaluator (Guo et al., 2025)	**87.1**	**67.5**	**69.6**	**74.7**

D.2 Adapted generative and vision-language models

This section presents the interface settings and results for adapted models.

Motion-to-text likelihood adapters. MotionGPT, MotionGPT3, MG-MotionLLM, and IRG-MotionLLM are tested through motion-to-text (M2T) adapters (Jiang et al., 2023; Zhu et al., 2025; Wu et al., 2025; Li et al., 2026). Given a motion, we compute the probability of each token in an existing caption and use the mean log-probability as its score. At each step, we condition on the caption's actual preceding tokens rather than tokens generated by the model. We include the end token but exclude prompt and padding tokens.

Vision-language forced-choice interface. Qwen-VL returns a categorical choice; its Δ uses the matched and unrelated-motion choice rates.

VLMs receive six uniformly spaced skeleton frames with root trajectories, ordered left to right. Prompts specify time order and request A/B, with randomized caption positions. Controls substitute a different motion ID within the same corpus–axis cell.

Results and analysis. Table 5 summarizes results for the two interfaces above.

MG-MotionLLM leads the likelihood adapters on structural scores, while IRG Stage-1 is near or below zero. Weak Random Content matching limits interpretation of Qwen-VL's structural scores.

Table 5: Adapted non-evaluator interfaces. (a) M2T models use length-normalized $\log p(c \mid m)$; (b) Qwen-VL uses six-keyframe forced choice. Metrics follow Table 2; Macro GPR averages three structural axes (null 50%). Panel (b) reports only acc for Random Content. In (a), bold marks best acc, Δ, and GPR, and underline marks second-best Δ and GPR. In (b), bold marks best acc and Δ per corpus, ranked before rounding.

(a) M2T Likelihood Adapters on HumanML3D

	Temporal Order			Mirror Reflection			Action Identity			Random Content			Macro
M2T likelihood adapter	acc	bias	$\Delta\uparrow$	acc	bias	$\Delta\uparrow$	acc	bias	$\Delta\uparrow$	acc	bias	$\Delta\uparrow$	GPR (%)↑
MotionGPT (Jiang et al., 2023)	0.902	0.894	0.081	0.730	0.514	<u>0.445</u>	0.764	0.679	0.265	0.621	0.488	0.261	62.9
MotionGPT3 (Zhu et al., 2025)	0.935	0.873	<u>0.492</u>	0.704	0.526	0.374	0.783	0.557	<u>0.511</u>	**0.767**	0.518	**0.516**	<u>68.9</u>
MG-MotionLLM (Wu et al., 2025)	**0.982**	0.960	**0.541**	**0.762**	0.514	**0.510**	**0.868**	0.557	**0.702**	0.747	0.514	<u>0.480</u>	**69.7**
IRG-MotionLLM (Stage-1) (Li et al., 2026)	0.953	0.955	-0.048	0.475	0.475	0.000	0.726	0.726	0.000	0.541	0.522	0.041	50.5

(b) Qwen-VL Keyframe Forced-Choice Adapters

		Temporal Order			Mirror Reflection			Action Identity			Random Content
Caption regime	Model	acc	bias	$\Delta\uparrow$	acc	bias	$\Delta\uparrow$	acc	bias	$\Delta\uparrow$	accuracy ↑
HumanML3D	Qwen3-VL-8B	0.620	0.567	0.123	0.513	0.480	**0.064**	0.616	0.554	0.140	**0.687**
	Qwen2-VL-7B	**0.753**	0.660	**0.275**	**0.547**	0.553	-0.015	**0.670**	0.554	**0.260**	0.680
SnapMoGen	Qwen3-VL-8B	**0.807**	0.747	**0.237**	**0.540**	0.493	**0.092**	**0.613**	0.540	**0.159**	**0.633**
	Qwen2-VL-7B	0.767	0.740	0.103	0.500	0.513	-0.027	0.607	0.573	0.078	0.520

E Structural hard-negative training and results

Neg Suite introduces structural hard negatives on the caption and motion sides, in addition to ordinary contrastive negatives, to strengthen supervision for fine-grained structural distinctions. We retain ordinary negatives and use captions or motions with similar content but different structure to provide additional supervision.

E.1 Losses and training settings

Caption-side negatives cover Temporal Order, Mirror Reflection, and Action Identity. The additional caption-pair loss for Action Identity and its combination with the pooled contrastive loss are:

$$\mathcal{L}_{\text{pair}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \text{softplus}\left(\frac{s(m_i, c_i') - s(m_i, c_i)}{\tau} \right), \qquad \mathcal{L} = \mathcal{L}_{\text{pool}} + 0.5\mathcal{L}_{\text{pair}}.$$

Here $\mathcal{A}$ indexes action pairs, c_i and c_i' are correct and action-edited captions, $\tau = 0.1$, and $\mathcal{L}_{\text{pool}}$ is the pooled contrastive loss.

All Neg Suite variants share the optimization settings below.

Setting	Value
Optimizer	AdamW
Learning Rate / Weight Decay	10^{-5} / 10^{-4}
Batch Size / Epochs	32 / 2
Temperature / Gradient-Norm Limit	0.1 / 1.0
Random Seed	42

E.2 Detailed intervention results

Pooled training increases mean diagonal interaction in the mirror audit: TMR from 0.094 to 0.297, SnapMoGen from 0.130 to 0.230.

NegHML selects the correct mirror caption at 0.600 accuracy and $\Delta = 0.127$, compared with 0.463 and -0.041 for the original evaluator. On the same 948 bidirectional pairs, diagonal accuracy rises from 48.05% to 71.04%. Motion-side accuracy is 61.92% under the original captions and 61.71% under the mirrored captions.

Table 6: Structural-hard-negative interventions. Metrics follow Table 2. Bold marks better acc and Δ within each base–finetuned pair; bias is not ranked. MedR is median retrieval rank (lower is better). All variants use caption negatives, only NegHML adds paired motion supervision, and ‡ marks separate Action Identity checkpoints (NegTMR-A and NegSnap-A), with other NegTMR/NegSnap results from pooled checkpoints.

	Temporal Order			Mirror Reflection			Action Identity			Random Content			Retrieval
Model	acc↑	bias	Δ ↑	acc↑	bias	Δ ↑	acc↑	bias	Δ ↑	acc↑	bias	Δ ↑	MedR↓
HumanML3D evaluator	0.624	0.392	0.381	0.463	0.484	-0.041	**0.858**	0.500	**0.717**	0.940	0.503	0.879	77
+NegHML (ours)	**0.732**	0.418	**0.540**	**0.600**	0.542	**0.127**	0.849	0.538	0.673	**0.958**	0.524	**0.912**	66
TMR	0.704	0.494	0.416	0.728	0.527	0.424	**0.887**	0.557	0.745	**0.955**	0.506	**0.910**	25
+NegTMR suite (ours)	**0.760**	0.483	**0.536**	**0.738**	0.537	**0.435**	**0.887**	0.547	**0.750**	0.953	0.511	0.903	24
SnapMoGen evaluator	0.965	0.518	0.927	0.728	0.519	0.435	**0.838**	0.529	**0.656**	0.998	0.500	0.995	1
+NegSnap suite (ours)	**0.970**	0.539	**0.935**	**0.742**	0.488	**0.496**	0.833	0.529	0.646	**0.999**	0.482	**0.998**	1

F Qualitative failure cases

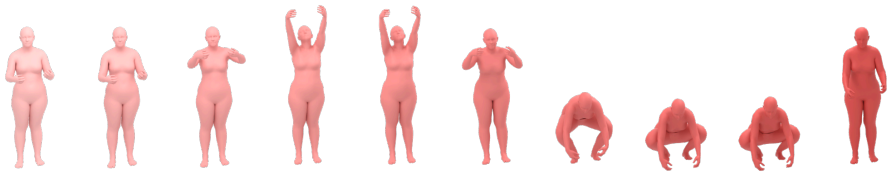
Here we provide more visualization cases where existing evaluators fail to detect the perturbations. Motions are fixed and captions are perturbed. Scores are negative raw $\text{sim}(m, c) - \text{sim}(m, c')$ margins. Each row shows ten uniformly sampled SMPL-H poses, earlier to later from left to right and light to dark.



Correct waving hands up and **down then walking around.**

Wrong **Walking around, then** waving hands up and **down.**

Correct - Wrong score: **-1.781**



Correct a person uses both of their hands to pick up **an** object above head **level, then crouches to put it down, then stands up straight.**

Wrong A person **stands up straight, then crouches to put down an object, then** uses both of their hands to pick up **the** object above head **level.**

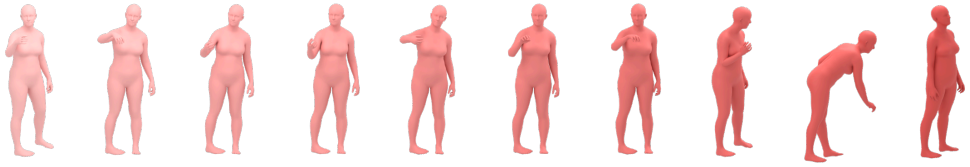
Correct - Wrong score: **-1.845**



Correct a man **seated lifts both hands in a "t-pose" then lowers them and then** stretches out his right **arm.**

Wrong A man stretches out his right **arm, then lowers both hands from a "t-pose", then lifts them.**

Correct - Wrong score: **-1.997**



Correct person rotates hand in air lots of times **then brings hand to right hip**

Wrong Person **brings hand to right hip, then** rotates hand in air lots of times

Correct - Wrong score: **-1.975**

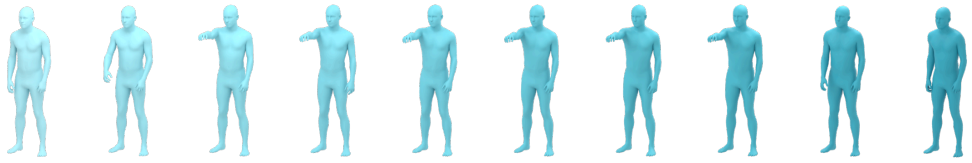
Earlier — Later

Figure 9: Temporal Order failures of the evaluators.



Correct the person was holding the **right** side of his head.
Wrong the person was holding the **left** side of his head.

Correct - Wrong score: **-0.370**



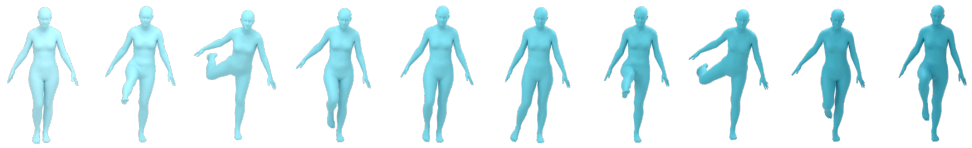
Correct this person lifts his **right** arm out front and points.
Wrong this person lifts his **left** arm out front and points.

Correct - Wrong score: **-0.177**



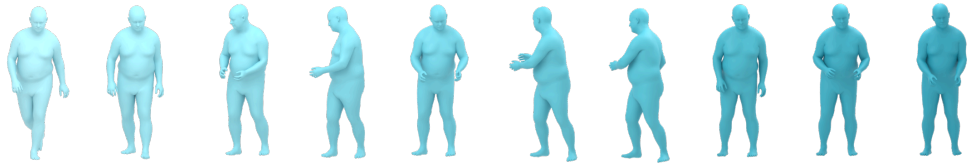
Correct person thrusts **right** leg forward in lunge three times and extends upper body forward three times at same time
Wrong person thrusts **left** leg forward in lunge three times and extends upper body forward three times at same time

Correct - Wrong score: **-0.208**



Correct a person lifts his **right** leg around in the air.
Wrong a person lifts his **left** leg around in the air.

Correct - Wrong score: **-0.189**



Correct a person walks forwards and swings their **right** arm backwards twice.
Wrong a person walks forwards and swings their **left** arm backwards twice.

Correct - Wrong score: **-0.225**

Earlier — Later

Figure 10: Mirror Reflection failures of the evaluators.



Correct a kneeling person crawls forward, turns around, crawls back, **stands** and turns to the left.
Wrong a kneeling person crawls forward, turns around, crawls back, **sits** and turns to the left.
Correct - Wrong score: **-0.158**



Correct a person **runs** forward and then stops
Wrong a person **walks** forward and then stops
Correct - Wrong score: **-0.910**



Correct the man is **sitting** then lays down.
Wrong the man is **standing** then lays down.
Correct - Wrong score: **-0.479**



Correct a man **runs** hurriedly in place before backstepping, dropping to a crouch, and duckwalking forward several steps with his arms bent in front of him at chest level.
Wrong a man **walks** hurriedly in place before backstepping, dropping to a crouch, and duckwalking forward several steps with his arms bent in front of him at chest level.
Correct - Wrong score: **-2.081**



Correct a person drops his arms, claps his hands together, then **walks** four steps to his right and sits down.
Wrong a person drops his arms, claps his hands together, then **runs** four steps to his right and sits down.
Correct - Wrong score: **-0.042**

Earlier — Later

Figure 11: Action Identity failures of the evaluators.