Event-T2M: Event-level Conditioning for Complex Text-to-Motion Synthesis

ICLR 2026

Seong-Eun Hong, JaeYoung Seon, JuYeong Hwang,
JongHwan Shin, HyeongYeop Kang

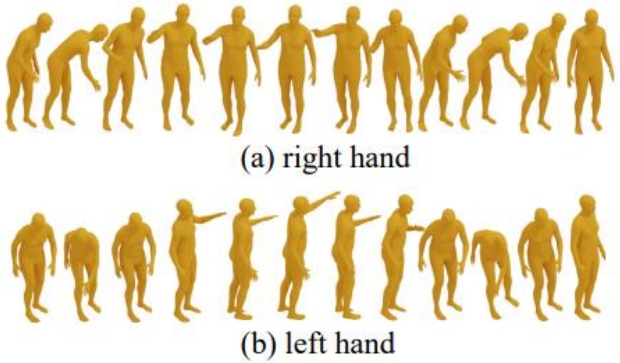
KOREA UNIVERSITY 1905

KYUNG HEE UNIVERSITY

Text-to-Motion Synthesis

- Previous methods
  - ParCo (ECCV 2024)
    - ✓ Problem: Body-part-specific prompts are not accurately reflected in the output
    - ✓ Solution: Modularize the model by creating separate networks for each body part
  - Light-T2M (AAAI 2025)
    - ✓ Problem: Existing models are not small enough for practical use
    - ✓ Solution: Improve performance while reducing model size by leveraging Mamba and CNN modules

"a man picks something with his [part description], shakes it, then puts it back."



[ParCo, 2024]

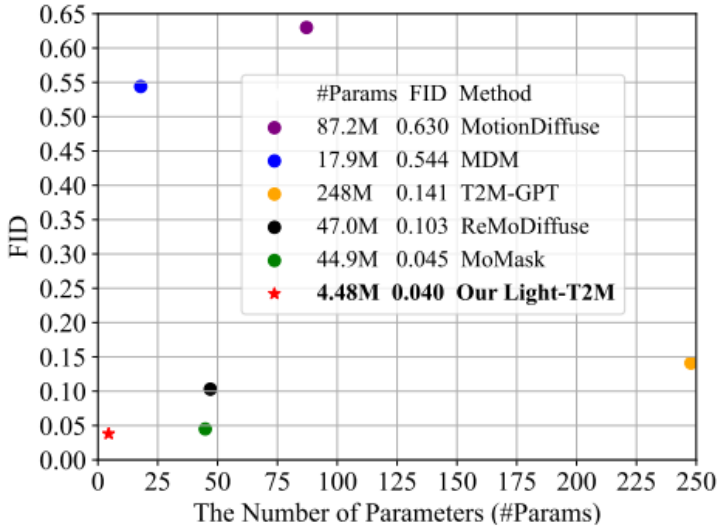
[Light-T2M, 2025]

Text-to-Motion Synthesis

- Previous methods
  - ParCo (ECCV 2024)
    - ✓ Problem: Body-part-specific prompts are not accurately reflected in the output
    - ✓ Solution: Modularize the model by creating separate networks for each body part
  - Light-T2M (AAAI 2025)
    - ✓ Problem: Existing models are not small enough for practical use
    - ✓ Solution: Improve performance while reducing model size by leveraging Mamba and CNN modules

"a man picks something with his [part description], shakes it, then puts it back."

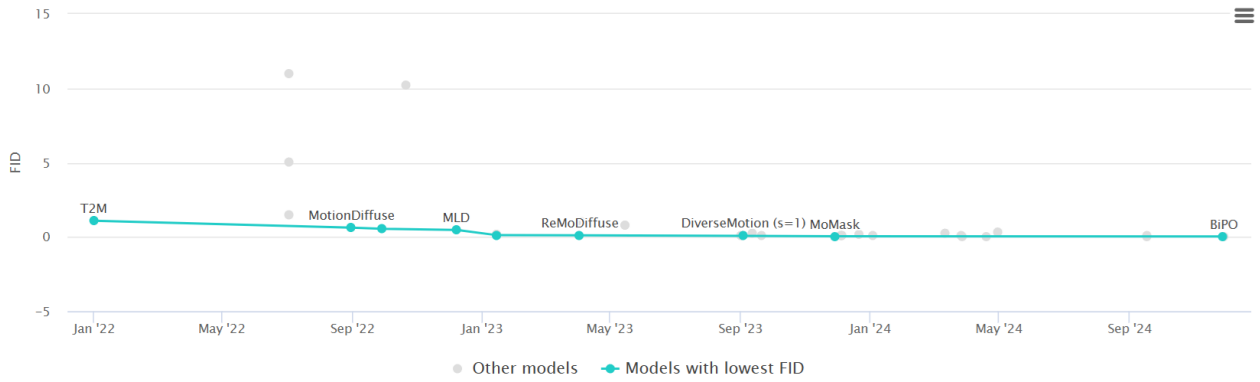
[ParCo, 2024]



[Light-T2M, 2025]

Text-to-Motion Synthesis

- Challenge
  - Evaluation methods dependent on existing benchmarks like HumanML3D and KIT-ML
    - ✓ Even papers claiming to generate complex motions well usually base their claims on strong benchmark results, particularly FID scores
    - ✓ The same is true for papers that claim state-of-the-art performance



Filter: untagged

Edit Leaderboard

Rank	Model	FID ↓	R Precision Top3	Diversity	Multimodality	Extra Training Data	Paper	Code	Result	Year	Tags
1	**BiPO**	0.030	0.809	9.556	1.374	×	BiPO: Bidirectional Partial Occlusion Network for Text-to-Motion Synthesis			2024	
2	**DisCoRD** (+MoMask)	0.032	0.809		1.288	×	DisCoRD: Discrete Tokens to Continuous Motion via Rectified Flow Decoding			2024	
3	**MoMask**	0.045	0.807		1.241	×	MoMask: Generative Masked Modeling of 3D Human Motions			2023	

Text-to-Motion Synthesis

- Challenge
  - Evaluation methods dependent on existing benchmarks like HumanML3D and KIT-ML
    - ✓ Even papers claiming to generate complex motions well usually base their claims on strong benchmark results, particularly FID scores
    - ✓ The same is true for papers that claim state-of-the-art performance

Can FID on benchmarks really represent the quality of all motions?

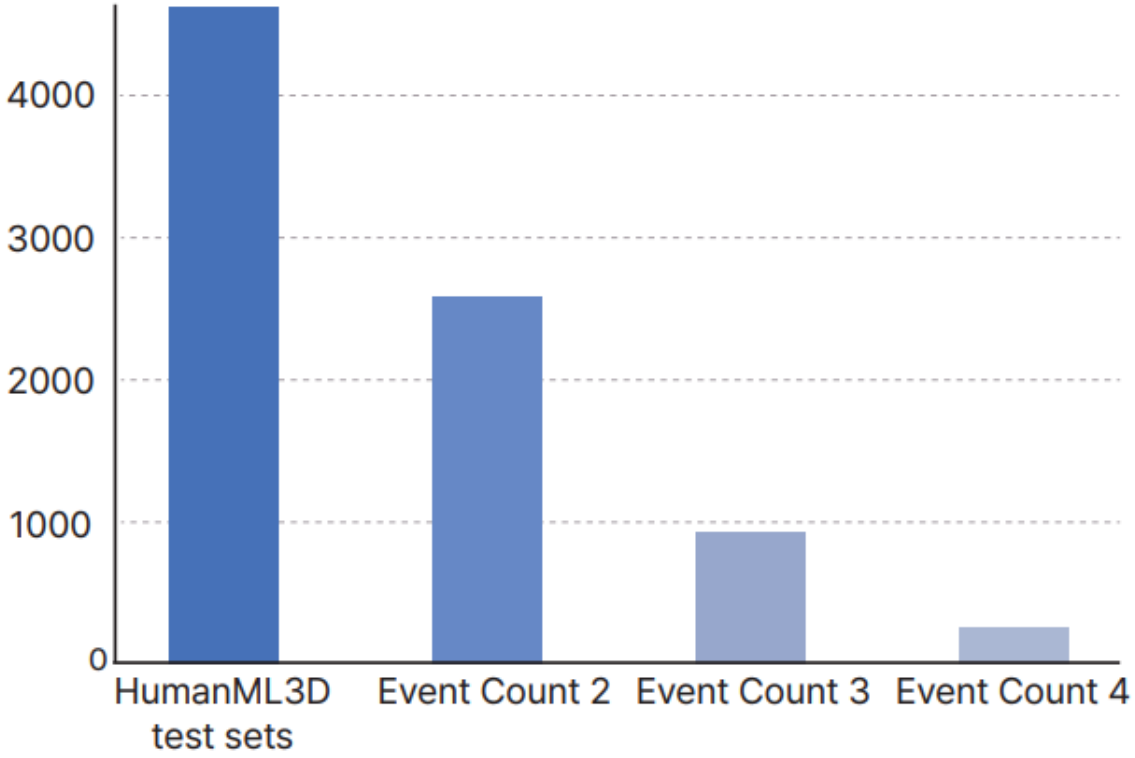
Filter: untagged

Edit Leaderboard

Rank	Model	FID ↓	R Precision Top3	Diversity	Multimodality	Extra Training Data	Paper	Code	Result	Year	Tags
1	BiPO	0.030	0.809	9.556	1.374	×	BiPO: Bidirectional Partial Occlusion Network for Text-to-Motion Synthesis			2024	
2	DisCoRD (+MoMask)	0.032	0.809		1.288	×	DisCoRD: Discrete Tokens to Continuous Motion via Rectified Flow Decoding			2024	
3	MoMask	0.045	0.807		1.241	×	MoMask: Generative Masked Modeling of 3D Human Motions			2023	

Problem setting

- Statistical analysis of text-to-motion synthesis benchmark
  - Using an LLM to take motion text as input and perform event-based statistical analysis
    - ✓ Gemini is used as the LLM
    - ✓ Conducted on the HumanML3D test set
    - ✓ Overwhelmingly, most motions contained only a single event



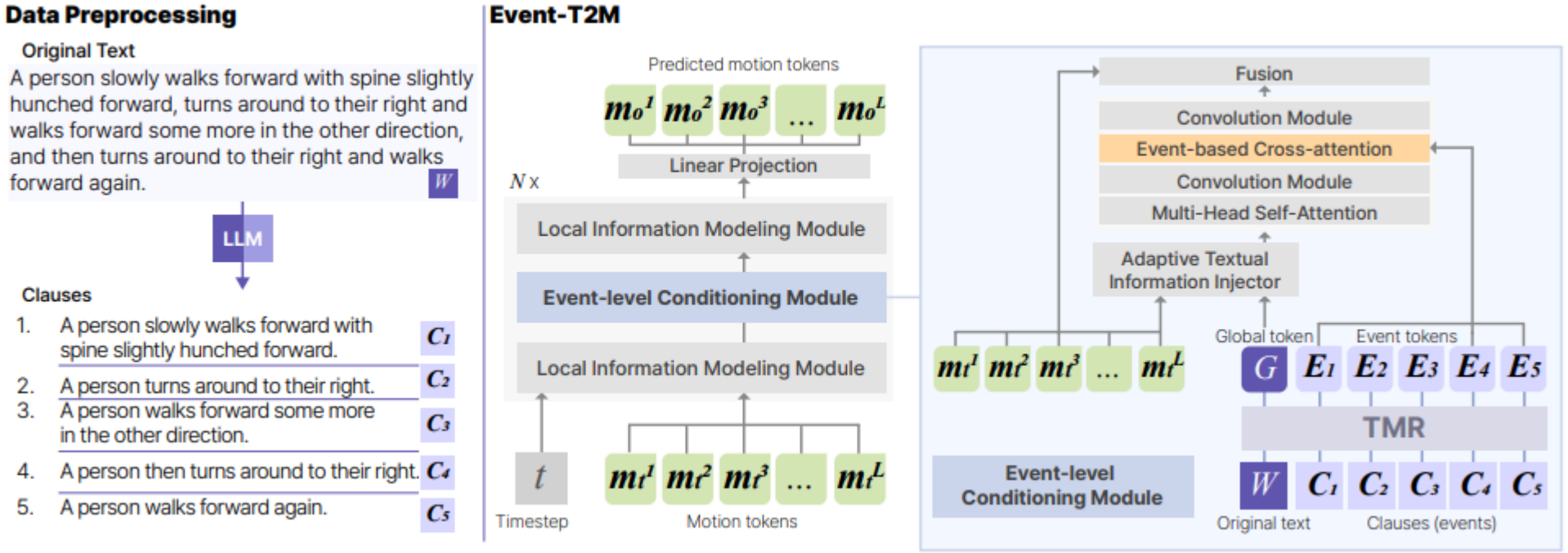
Problem setting

- Statistical analysis of text-to-motion synthesis benchmark
  - FID analysis of existing SOTA models on test samples with 4+ events
    - ✓ AttT2M 0.112->1.077 (+861.6%)
    - ✓ GraphMotion 0.116->0.857 (+638.8%)
    - ✓ MoMask 0.045->0.418 (+828.9%)
    - ✓ Light-T2M 0.040->0.627 (+1467.5%)
    - ✓ MoGenTS 0.033->0.423 (+1181.8%)

Existing models may generate simple motions well, but they struggle to produce complex motions effectively!

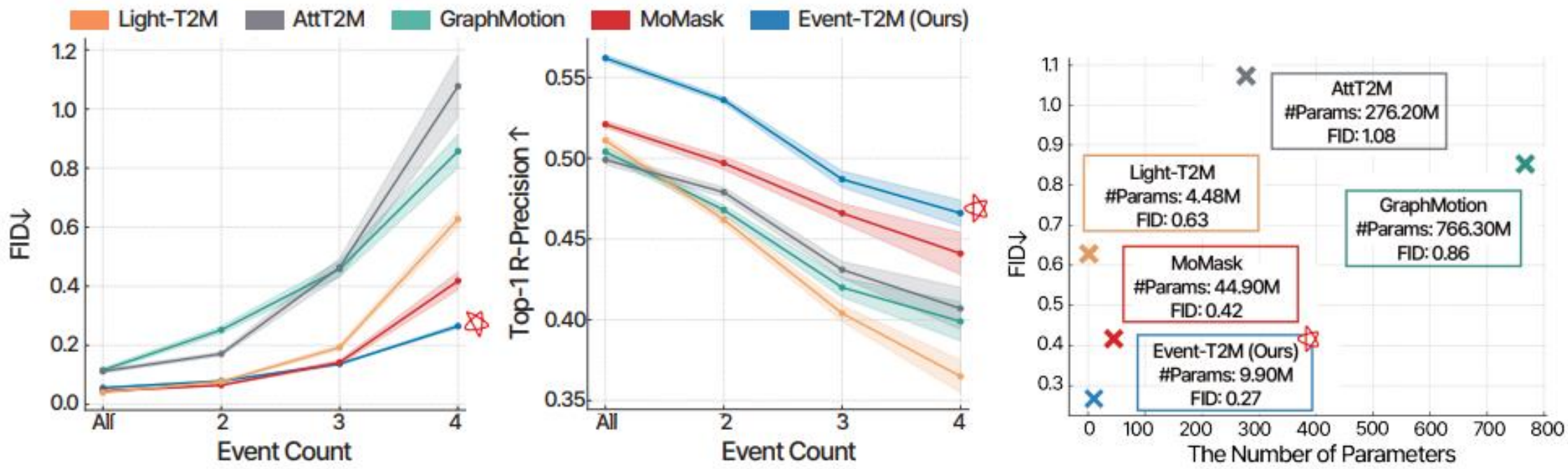
Methodology

- Our goal
  - Building a model that generates not only simple motions well, but also complex motions effectively
- Proposed methods – EventT2M
  - Instead of feeding arbitrary text directly into CLIP for token- or vocabulary-level conditioning, use an LLM to segment the text and perform event-level conditioning
  - Instead of CLIP, use TMR, which contains motion-related knowledge, for conditioning



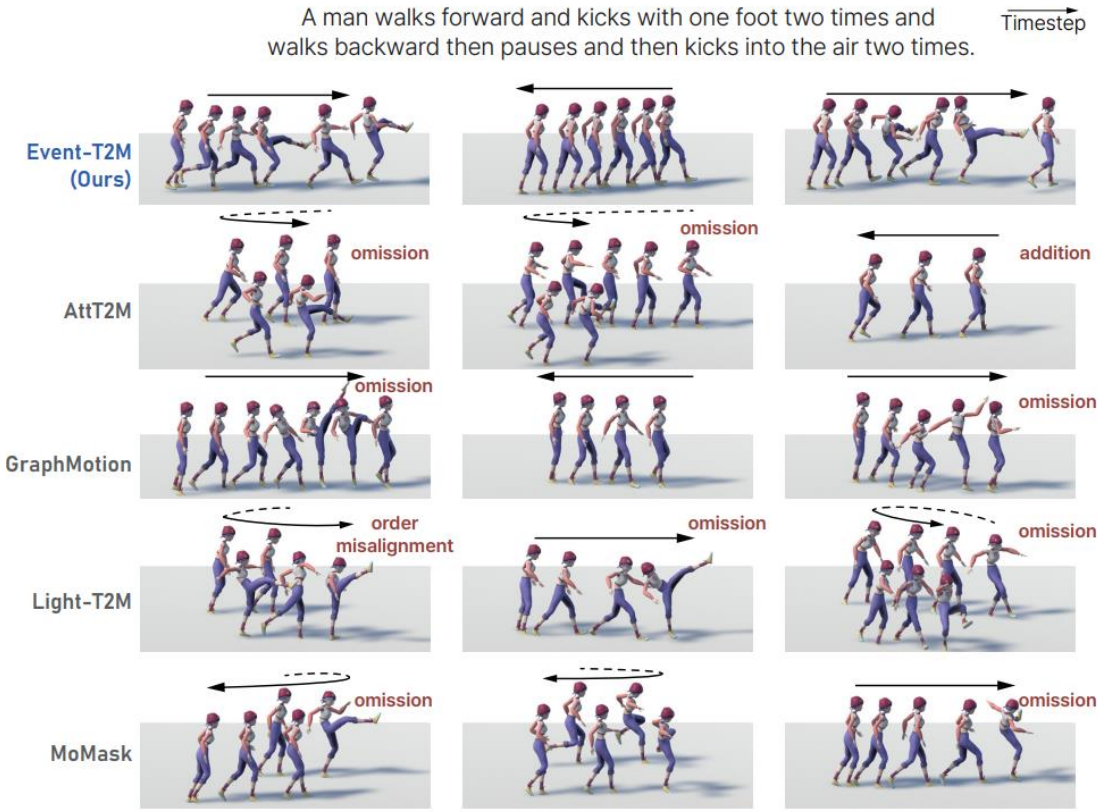
Quantitative results

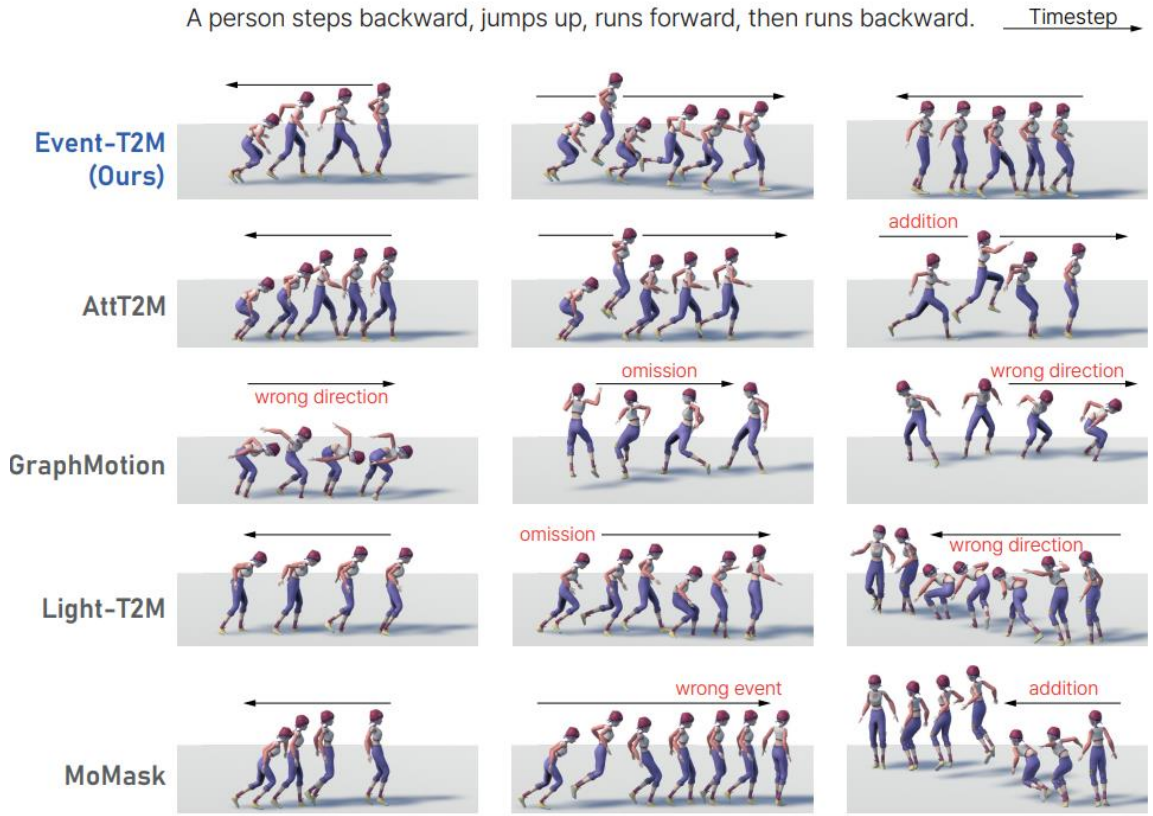
- Analysis of experimental results
  - While other models exhibit a sharp deterioration in performance, such as FID and R-precision, as the number of events increases, our model, EventT2M, shows only a modest drop in performance
  - Its smaller model size also makes it more cost-efficient



Qualitative results

- Analysis of visualized generated motions with four or more events
  - In qualitative examples of complex motion generation, other models often omit motions, add unintended ones, or produce incorrect movements, whereas our model generates them more accurately
  - The user study also showed an overwhelmingly strong qualitative preference for our model





Ablation study

- Ablation analysis of the text encoder and event-level conditioning
  - This demonstrates that TMR is used not merely because it contains motion-related knowledge, but because it is better suited for complex motion generation; in contrast, CLIP is actually more effective for simple motion generation
  - This demonstrates that explicit event-level conditioning is more effective than token-level conditioning, which only captures the meaning of individual vocabulary tokens

Text Encoder	Condition	Methods	R-Precision ↑			FID ↓	MM-Dist ↓	MModality ↑
			Top-1 ↑	Top-2 ↑	Top-3 ↑			
TMR	2	Event-T2M (Token-level)	$0.521^{\pm.003}$	$0.718^{\pm.002}$	$0.815^{\pm.002}$	$0.082^{\pm.003}$	$2.915^{\pm.008}$	$\mathbf{0.999}^{\pm.032}$
		Event-T2M (Event-level)	$\mathbf{0.536}^{\pm.002}$	$\mathbf{0.732}^{\pm.002}$	$\mathbf{0.824}^{\pm.002}$	$\mathbf{0.079}^{\pm.003}$	$\mathbf{2.836}^{\pm.006}$	$0.976^{\pm.043}$
	3	Event-T2M (Token-level)	$0.463^{\pm.005}$	$0.664^{\pm.005}$	$0.773^{\pm.003}$	$0.162^{\pm.006}$	$3.031^{\pm.009}$	$\mathbf{1.035}^{\pm.045}$
		Event-T2M (Event-level)	$\mathbf{0.487}^{\pm.005}$	$\mathbf{0.687}^{\pm.004}$	$\mathbf{0.790}^{\pm.004}$	$\mathbf{0.137}^{\pm.003}$	$\mathbf{2.928}^{\pm.010}$	$1.010^{\pm.029}$
	4	Event-T2M (Token-level)	$0.440^{\pm.011}$	$0.635^{\pm.010}$	$0.740^{\pm.009}$	$0.355^{\pm.011}$	$3.168^{\pm.016}$	$\mathbf{1.141}^{\pm.026}$
		Event-T2M (Event-level)	$\mathbf{0.466}^{\pm.008}$	$\mathbf{0.660}^{\pm.008}$	$\mathbf{0.767}^{\pm.007}$	$\mathbf{0.265}^{\pm.007}$	$\mathbf{3.063}^{\pm.015}$	$1.039^{\pm.028}$
CLIP	2	Event-T2M (Token-level)	$0.474^{\pm.003}$	$0.664^{\pm.003}$	$0.767^{\pm.003}$	$0.153^{\pm.004}$	$3.149^{\pm.010}$	$\mathbf{1.875}^{\pm.057}$
		Event-T2M (Event-level)	$\mathbf{0.494}^{\pm.003}$	$\mathbf{0.681}^{\pm.003}$	$\mathbf{0.779}^{\pm.003}$	$\mathbf{0.052}^{\pm.002}$	$\mathbf{3.079}^{\pm.010}$	$1.577^{\pm.060}$
	3	Event-T2M (Token-level)	$0.423^{\pm.006}$	$\mathbf{0.618}^{\pm.005}$	$0.728^{\pm.004}$	$0.206^{\pm.008}$	$3.254^{\pm.011}$	$\mathbf{1.905}^{\pm.056}$
		Event-T2M (Event-level)	$\mathbf{0.423}^{\pm.005}$	$\mathbf{0.618}^{\pm.005}$	$\mathbf{0.729}^{\pm.005}$	$\mathbf{0.141}^{\pm.004}$	$\mathbf{3.245}^{\pm.015}$	$1.627^{\pm.052}$
	4	Event-T2M (Token-level)	$\mathbf{0.399}^{\pm.012}$	$\mathbf{0.597}^{\pm.010}$	$\mathbf{0.709}^{\pm.010}$	$0.468^{\pm.021}$	$\mathbf{3.339}^{\pm.032}$	$\mathbf{1.991}^{\pm.060}$
		Event-T2M (Event-level)	$0.374^{\pm.010}$	$0.578^{\pm.007}$	$0.690^{\pm.007}$	$\mathbf{0.425}^{\pm.022}$	$3.467^{\pm.022}$	$1.674^{\pm.059}$

Conclusion

- Benchmark-based evaluation, particularly on HumanML3D and KIT-ML, is heavily biased toward simple motions and does not fully reflect performance on complex motion generation
- When evaluated on test samples with four or more events, existing SOTA models show a sharp degradation in FID and R-precision, indicating limited robustness to complex motions
- By leveraging TMR and explicit event-level conditioning, EventT2M generates both simple and complex motions more effectively while remaining compact and cost-efficient
- Qualitative examples and user study results further confirm that EventT2M produces more faithful complex motions and is overwhelmingly preferred over prior models

Conclusion

- Benchmark-based evaluation, particularly on HumanML3D and KIT-ML, is heavily biased toward simple motions and does not fully reflect performance on complex motion generation
- When evaluated on test samples with four or more events, existing SOTA models show a sharp degradation in FID and R-precision, indicating limited robustness to complex motions
- By leveraging TMR and explicit event-level conditioning, EventT2M generates both simple and complex motions more effectively while remaining compact and cost-efficient
- Qualitative examples and user study results further confirm that EventT2M produces more faithful complex motions and is overwhelmingly preferred over prior models

Conclusion

- Benchmark-based evaluation, particularly on HumanML3D and KIT-ML, is heavily biased toward simple motions and does not fully reflect performance on complex motion generation
- When evaluated on test samples with four or more events, existing SOTA models show a sharp degradation in FID and R-precision, indicating limited robustness to complex motions
- By leveraging TMR and explicit event-level conditioning, EventT2M generates both simple and complex motions more effectively while remaining compact and cost-efficient
- Qualitative examples and user study results further confirm that EventT2M produces more faithful complex motions and is overwhelmingly preferred over prior models

Thank you