Past as State, Present as Attention:

Persistent-State Blockwise Flow Matching for Long-Horizon Co-Speech Motion Generation

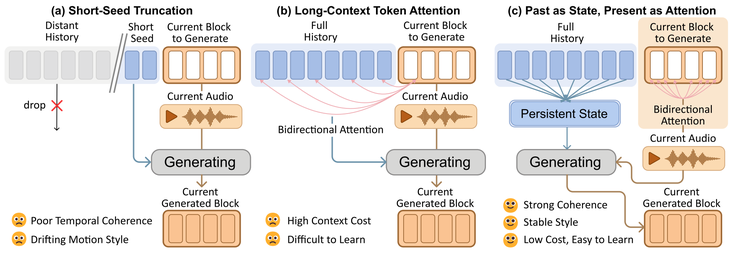
Yi Liu[1,3] Xiangyue Zhang[2] Jianfang Li[3]* Jia Ma[1] Yisheng He[3] Jianqiang Ren[3]

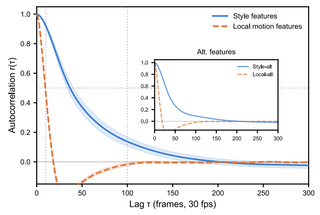
[1] Tongji University [2] The University of Tokyo [3] Alibaba Group * Corresponding author

NeurIPS 2026

Motivation and Key Insight

Short seeds lose speaker style. Full-history attention is costly and can underuse history. PASPA separates slow style from fast motion to preserve long-range consistency and local articulation.



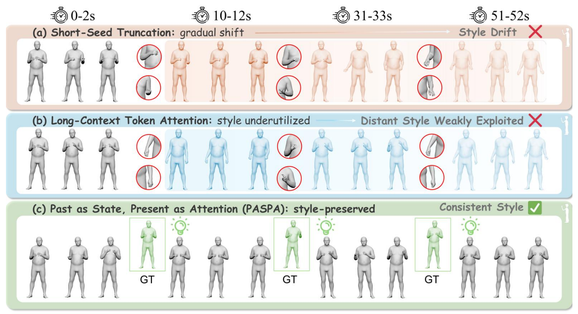


Past: slowly varying gesture amplitude and spatial preference.

Present: fast local dynamics and fine articulation.

Use a persistent state for history and full attention within each block.

Long-Horizon Style Preservation



Same speech and speaker: PASPA retains the hands-near-chest style for 52 s. Short seeds drift, while token attention preserves style inconsistently.

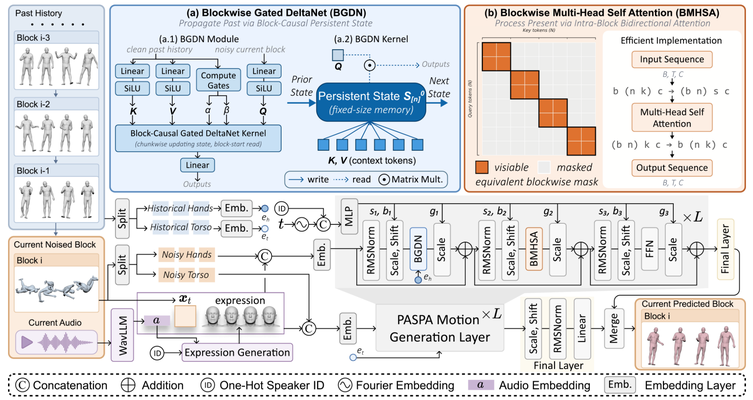
Long-Rollout Quality: Final 10-Second Windows

Method	[0, 10]s		[20, 30]s		[50, 60]s	
	FGD↓	Div↑	FGD↓	Div↑	FGD↓	Div↑
Short Seed	4.731	6.581	5.886	6.033	6.721	6.081
Token Attention	**4.448**	6.289	5.388	6.510	5.771	6.427
PASPA	**4.448**	**7.217**	**5.119**	**7.133**	**5.745**	**7.138**

Across 10 / 30 / 60 s rollouts, PASPA has the best FGD and diversity.

Method: Past as State, Present as Attention

BGDN reads a frozen block-start state to preserve past style. BMHSA models fine dynamics within the current block.



$$\mathbf{S}_i = \mathbf{S}_{i-1}\left[\alpha_i(\mathbf{I} - \beta_i \mathbf{k}_i\mathbf{k}_i^\top)\right] + \beta_i \mathbf{v}_i\mathbf{k}_i^\top$$

$$\mathbf{O}_{[n]}^{\mathrm{BGDN}} = \mathbf{Q}_{[n]}(\mathbf{S}_{[n]}^0)^\top, \qquad \mathbf{S}_{[n]}^0 = \mathbf{S}_{[n-1]}^K$$

Training: Blockwise Flow Matching

Denoise all blocks in parallel using teacher-forced clean history, varied noise levels and variable-length contexts.

$$\mathbf{x}_{t_n}^n = (1 - t_n)\mathbf{x}_0^n + t_n \mathbf{z}^n, \quad \mathbf{z}^n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$t_1 \le t_2 \le \cdots \le t_N, \qquad \hat{\mathbf{x}}_0 = F_\theta(\mathbf{x}_\mathbf{t}, \mathbf{x}_0, \mathbf{t}, \mathbf{a})$$

$$\mathcal{L} = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} \ell_n + \mathcal{L}_{\mathrm{bd}}\right], \qquad \mathcal{L}_{\mathrm{geom}}^n = \mathcal{L}_{\mathrm{joint}}^n + \mathcal{L}_{\mathrm{struct}}^n$$

$$\ell_n = e^{-u(t_n)}\mathcal{L}_{\mathrm{rot}}^n + \mathrm{sg}(e^{-u(t_n)})\mathcal{L}_{\mathrm{geom}}^n + u(t_n)$$

Learned weights balance losses across noise levels. Boundary velocity and acceleration losses preserve cross-block continuity.

Inference: Cached Rollout and Hybrid CFG

Decode each block using a frozen history state. Prefill once with the completed clean block, then reuse the updated state.

$$\hat{\mathbf{x}}_0^{\mathrm{guided}} = f_\varnothing + w_c(f_c - f_\varnothing) + w_h(f_h - f_\varnothing) + w_c w_h(f_{ch} - f_c - f_h + f_\varnothing)$$

Four predictions separate speech and history effects. The interaction term jointly controls speech alignment and dependence on past motion.

Main Results: BEAT2 and SHOW

BEAT2 · Single-Speaker and All-Speaker Evaluation

Method	1 Speaker				All Speakers			
	FGD↓	BC↑	Div↑	MSE↓	FGD↓	BC↑	Div↑	MSE↓
CaMN [44]	6.64	0.677	10.86	–	5.12	0.200	5.58	–
DSG [81]	8.81	0.724	11.49	–	11.74	0.734	11.12	–
EMAGE [45]	5.51	0.772	13.06	7.680	6.92	0.284	6.06	6.908
DiffSHEG [11]	8.99	0.714	11.91	7.665	–	–	–	–
MambaTalk [79]	5.37	0.781	13.05	6.289	–	–	–	–
HoloGest [13]	5.34	**0.795**	**14.15**	–	6.46	**0.803**	**13.53**	–
RAG-GESTURE [54]	8.08	0.734	11.97	–	4.87	0.514	9.94	–
SemTalk [89]	4.28	0.777	12.91	6.153	–	–	–	–
GestureLSM [48]	4.09	0.714	13.24	–	–	–	–	–
GlobalDiff [92]	4.78	0.705	13.13	6.330	2.63	0.404	8.24	4.144
Ours	**3.71**	0.746	13.38	**5.768**	**2.33**	0.645	9.42	**3.793**

Best FGD and face MSE in both settings.

PASPA improves every reported metric over GlobalDiff.

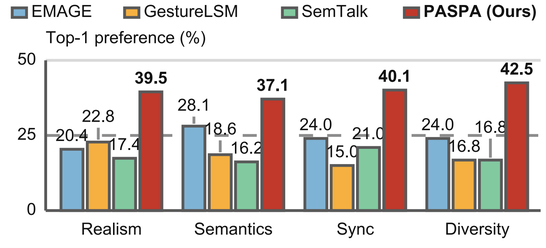
FGD ×10; MSE ×10^8. Bold / underline: best / second best.

SHOW · Cross-Dataset Evaluation

Method	FGD↓	BA↑	Div↑
GT	–	0.867	0.834
LS3DCG [24]	0.00478	**0.947**	0.645
TalkSHOW [85]	0.00328	0.872	0.711
DiffSHEG [11]	0.00271	0.902	0.810
GlobalDiff [92]	0.00198	0.870	**1.180**
PASPA	**0.00111**	0.876	1.043

0.00111 FGD: 43.9% lower than GlobalDiff. Speech–motion alignment also improves, using the same training recipe without dataset-specific tuning.

User Study: Preference across Four Criteria



39.8% overall top-1

Highest preference in all four criteria.

Blind ranking study:
18 participants
167 ranking groups
20-second clips

Ablation: Why the Hybrid Works

State + attention yields 5.216 FGD, versus 5.710 with full-history attention (FGD ×10).

Past motion benefits from a compact state; the current block needs dense local interaction.

Project

ly403.github.io/PASPA/