FloodDiffusion: Tailored Diffusion Forcing for Streaming Motion Generation

Supplementary Material

This supplemental document contains four sections:

- Video and Codes (Section A).
- Detailed Proofs of Theoretical Results (Section B).
- Baseline Implementation Details (Section C).
- Hyper-parameter Search (Section D).
- Details of User Study (Section E).

A. Video and Codes

We provide a comprehensive HTML that contains separate videos to demonstrate the performance of our system, including:

- Results for streaming human motion generation.
- Influence on the position to giving the text prompt.
- Stop the motion via neural language command.
- Compare with non-streaming methods.
- Compare with other streaming methods.
- Ablation study for the bi-directional attention and time scheduler.

B. Detailed Proofs of Theoretical Results

Proposition B.1 (Vectorized Conditional Dynamics)**.** *This proposition restates Proposition 3.2 in the main text. Under the vectorized time schedule, the conditional vector field and score function of the Gaussian path*

$$p_t(\mathbf{x} \mid \mathbf{z}) = \prod_{k=0}^{K-1} \mathcal{N}\big(x^k; \alpha_t^k z^k, (\beta_t^k)^2 \mathbf{I}\big) \tag{27}$$

are given by

$$u_t(\mathbf{x} \mid \mathbf{z}) = \left(\dot{\boldsymbol{\alpha}}_t - \frac{\dot{\boldsymbol{\beta}}_t}{\boldsymbol{\beta}_t} \odot \boldsymbol{\alpha}_t\right) \odot \mathbf{z} + \left(\frac{\dot{\boldsymbol{\beta}}_t}{\boldsymbol{\beta}_t}\right) \odot \mathbf{x} \tag{28}$$

$$s_t(\mathbf{x} \mid \mathbf{z}) = -\frac{(\mathbf{x} - \boldsymbol{\alpha}_t \odot \mathbf{z})}{\boldsymbol{\beta}_t^2} \tag{29}$$

Proof. Since each dimension is independent in the Gaussian distribution, we have

$$p_t(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\big(\mathbf{x};\, \boldsymbol{\alpha}_t \odot \mathbf{z}, \operatorname{diag}(\boldsymbol{\beta}_t^2)\big) = \prod_{i=1}^{K} p_t^i(x^i \mid z^i) \tag{30}$$

where $p_t^i(x^i \mid z^i)$ denotes the marginal density of the i-th coordinate.

For a Gaussian $\mathcal{N}(x; \mu, \Sigma)$, the score is

$$\nabla_x \log p(x) = -\Sigma^{-1}(x - \mu) \tag{31}$$

Applying this to our diagonal-covariance case gives

$$s_t(\mathbf{x} \mid \mathbf{z}) = \nabla_{\mathbf{x}} \log p_t(\mathbf{x} \mid \mathbf{z}) \tag{32}$$

$$= -\operatorname{diag}(\boldsymbol{\beta}_t^{-2})\big(\mathbf{x} - \boldsymbol{\alpha}_t \odot \mathbf{z}\big) \tag{33}$$

$$= -\frac{\mathbf{x} - \boldsymbol{\alpha}_t \odot \mathbf{z}}{\boldsymbol{\beta}_t^2} \tag{34}$$

which matches the claimed expression.

We now verify that u_t defines a valid probability flow field by checking the continuity equation

$$\partial_t p_t(\mathbf{x}) + \nabla_{\mathbf{x}} \cdot \big(p_t(\mathbf{x})\, u_t(\mathbf{x})\big) = 0 \tag{35}$$

Using the factorization $p_t(\mathbf{x}) = \prod_{i=1}^{K} p_t^i(x^i)$ and writing $u_t(\mathbf{x}) = (u_t^1(x^1), \ldots, u_t^K(x^K))$, we obtain

$$\partial_t p_t(\mathbf{x}) = \partial_t \Big(\prod_{i=1}^{K} p_t^i(x^i)\Big) \tag{36}$$

$$= \sum_{i=1}^{K} \left[\Big(\partial_t p_t^i(x^i)\Big) \prod_{j \neq i} p_t^j(x^j)\right] \tag{37}$$

$$\nabla_{\mathbf{x}} \cdot \big(p_t(\mathbf{x})\, u_t(\mathbf{x})\big) \tag{38}$$

$$= \sum_{i=1}^{K} \partial_{x^i} \Big(p_t(\mathbf{x})\, u_t^i(x^i)\Big) \tag{39}$$

$$= \sum_{i=1}^{K} \partial_{x^i} \left(\Big[\prod_{j=1}^{K} p_t^j(x^j)\Big] u_t^i(x^i)\right) \tag{40}$$

$$= \sum_{i=1}^{K} \left[\Big(\partial_{x^i} p_t^i(x^i)\Big) u_t^i(x^i)\right] \prod_{j \neq i} p_t^j(x^j) \tag{41}$$

$$+ \sum_{i=1}^{K} \left[p_t^i(x^i)\, \partial_{x^i} u_t^i(x^i)\right] \prod_{j \neq i} p_t^j(x^j) \tag{42}$$

Therefore,

$$\partial_t p_t(\mathbf{x}) + \nabla_{\mathbf{x}} \cdot \big(p_t(\mathbf{x})\, u_t(\mathbf{x})\big) \tag{43}$$

$$= \sum_{i=1}^{K} \left[\prod_{j \neq i} p_t^j(x^j)\right] \Big(\partial_t p_t^i(x^i) + \partial_{x^i}\big(p_t^i(x^i)\, u_t^i(x^i)\big)\Big) \tag{44}$$

The term in parentheses is exactly the one-dimensional continuity equation for the i-th coordinate, which holds by construction of the scalar Gaussian flow.

Hence each summand is zero, and the whole sum vanishes, proving that u_t is a valid vector field for the factorized Gaussian path. □

Theorem B.2 (Conditional Generation)**.** *This theorem restates Theorem 3.4 in the main text. Consider the SDE*

$$\mathbf{X}_0 \sim p_{init}, \tag{45}$$

$$\mathrm{d}\mathbf{X}_t = \left[u_t(\mathbf{X}_t, \mathbf{c}) + \frac{\boldsymbol{\sigma}_t^2}{2} s_t(\mathbf{X}_t, \mathbf{c})\right] \mathrm{d}t + \sigma_t \mathrm{d}\mathbf{W}_t \tag{46}$$

where u_t and s_t are respectively the marginal vector field and score of the path $p_t(\mathbf{x}, \mathbf{c})$. Then the marginal law of $\mathbf{X}_t$ satisfies $\mathbf{X}_t \sim p_t(\cdot \mid \mathbf{c})$ for all t, and in particular $\mathbf{X}_T \sim p_{data}(\cdot \mid \mathbf{c})$.

Proof. We verify the claim by appealing to the Fokker–Planck equation. We first work in the general matrix-valued setting, then specialize to the diagonal case.

General case. Consider a generic SDE of the form

$$\mathrm{d}\mathbf{x}_t = u_t^{fp}(\mathbf{x}_t)\,\mathrm{d}t + \sigma_t\,\mathrm{d}\mathbf{W}_t \tag{47}$$

with initial condition $\mathbf{x}_0 \sim p_{\text{init}}$, where σ_t is a diffusion matrix (so that $\sigma_t\sigma_t^\top$ is the covariance). The Fokker–Planck equation states that the marginal density p_t evolves as

$$\partial_t p_t(\mathbf{x}) = -\nabla_{\mathbf{x}} \cdot \big(p_t(\mathbf{x}) u_t^{fp}(\mathbf{x})\big) + \frac{1}{2}\nabla_{\mathbf{x}} \cdot \Big(\sigma_t\sigma_t^\top \nabla_{\mathbf{x}} p_t(\mathbf{x})\Big) \tag{48}$$

To match the desired marginal $p_t(\cdot \mid \mathbf{c})$, we decompose the drift by adding and subtracting a score-correction term (we omit $\mathbf{c}$ for brevity):

$$u_t(\mathbf{x}) = u_t(\mathbf{x}) + \frac{\sigma_t\sigma_t^\top}{2} s_t(\mathbf{x}) - \frac{\sigma_t\sigma_t^\top}{2} s_t(\mathbf{x}) \tag{49}$$

Substituting this into the FP equation yields

$$\partial_t p_t(\mathbf{x}) = -\nabla_{\mathbf{x}} \cdot \big(p_t u_t\big) \tag{50}$$

$$= -\nabla_{\mathbf{x}} \cdot \Big(p_t\big(u_t + \tfrac{\sigma_t\sigma_t^\top}{2} s_t\big)\Big) \tag{51}$$

$$+ \nabla_{\mathbf{x}} \cdot \Big(p_t \tfrac{\sigma_t\sigma_t^\top}{2} s_t\Big) \tag{52}$$

Applying the identity $s_t = \nabla_{\mathbf{x}} \log p_t$, which gives $p_t \nabla_{\mathbf{x}} \log p_t = \nabla_{\mathbf{x}} p_t$, we find

$$\nabla_{\mathbf{x}} \cdot \Big(p_t \tfrac{\sigma_t\sigma_t^\top}{2} s_t\Big) = \frac{1}{2}\nabla_{\mathbf{x}} \cdot \Big(\sigma_t\sigma_t^\top \nabla_{\mathbf{x}} p_t\Big) \tag{53}$$

This precisely reproduces diffusion term in the FP equation, leaving

$$\partial_t p_t(\mathbf{x}) = -\nabla_{\mathbf{x}} \cdot \Big(p_t(\mathbf{x})\big(u_t(\mathbf{x}) + \tfrac{\sigma_t\sigma_t^\top}{2} s_t(\mathbf{x})\big)\Big) \tag{54}$$

$$+ \frac{1}{2}\nabla_{\mathbf{x}} \cdot \Big(\sigma_t\sigma_t^\top \nabla_{\mathbf{x}} p_t(\mathbf{x})\Big) \tag{55}$$

Specialization to diagonal diffusion. We now restrict to the diagonal case where $\sigma_t = \mathrm{diag}(\boldsymbol{\sigma}_t)$ and hence $\sigma_t\sigma_t^\top = \mathrm{diag}(\boldsymbol{\sigma}_t^2)$, with $\boldsymbol{\sigma}_t^2$ denoting the element-wise square of the noise standard-deviation vector. In this case, the drift correction simplifies to

$$u_t(\mathbf{x}, \mathbf{c}) + \tfrac{\sigma_t\sigma_t^\top}{2} s_t(\mathbf{x}, \mathbf{c}) = u_t(\mathbf{x}, \mathbf{c}) + \tfrac{\boldsymbol{\sigma}_t^2}{2} s_t(\mathbf{x}, \mathbf{c}) \tag{56}$$

where all operations are now coordinate-wise.

Therefore, for any diagonal diffusion covariance $\Sigma_t = \mathrm{diag}(\boldsymbol{\sigma}_t^2)$, the SDE (14) with drift $u_t^{fp} = u_t + \frac{\boldsymbol{\sigma}_t^2}{2} s_t$ produces the target marginal distribution. Since u_t and s_t are defined (Definition 3.3) to generate the marginal path $p_t(\cdot \mid \mathbf{c})$, we conclude that $\mathbf{x}_t \sim p_t(\cdot \mid \mathbf{c})$ for all $t \in [0, T]$, and in particular $\mathbf{x}_T \sim p_{\text{data}}(\cdot \mid \mathbf{c})$. □

Lemma B.3 (Schedule Saturation)**.** *This lemma restates Lemma 3.6 in the main text. Under the schedule*

$$\alpha_t^k = \mathrm{clamp}(t - \tfrac{k}{n_s}, 0, 1), \qquad \beta_t^k = 1 - \alpha_t^k \tag{57}$$

for any t we have
1. *If $k < m(t) = \lceil (t-1)n_s \rceil$, then $\alpha_t^k = 1$ and $\beta_t^k = 0$;*
2. *If $k \geq n(t) = \lceil t n_s \rceil$, then $\alpha_t^k = 0$ and $\beta_t^k = 1$.*

Proof. Fix t and k. By definition

$$\alpha_t^k = \mathrm{clamp}\big(t - \tfrac{k}{n_s}, 0, 1\big) \tag{58}$$

If $k < m(t) = \lceil (t-1)n_s \rceil$, then $k/n_s < t - 1$, hence $t - k/n_s > 1$ and $\alpha_t^k = 1$. Consequently $\beta_t^k = 1 - \alpha_t^k = 0$.

Conversely, if $k \geq n(t) = \lceil t n_s \rceil$, then $k/n_s \geq t$ so that $t - k/n_s \leq 0$ and thus $\alpha_t^k = 0$ and $\beta_t^k = 1$. This establishes both claims. □

Theorem B.4 (Streaming Locality)**.** *This theorem restates Theorem 3.8 in the main text. Under Assumption 3.7 and the triangular schedule (19)–(20), the velocity field factorizes as*

$$u_t(\mathbf{X}_t, \mathbf{c}^{0:K}) = \begin{bmatrix} \mathbf{0}^{0:m(t)} \\ u_t^{m(t):n(t)}(\mathbf{X}_t^{0:n(t)}, \mathbf{c}^{0:n(t)}) \\ \mathbf{0}^{n(t):K} \end{bmatrix} \tag{59}$$

Proof. We first observe that the conditional velocity u_t, noise ϵ_t, data prediction z_t, and score s_t can all be expressed as affine combinations of $\mathbf{z}$ and $\mathbf{x}$:

$$u_t(\mathbf{x} \mid \mathbf{z}) = \left(\dot{\boldsymbol{\alpha}}_t - \frac{\dot{\boldsymbol{\beta}}_t}{\boldsymbol{\beta}_t} \odot \boldsymbol{\alpha}_t\right) \odot \mathbf{z} + \left(\frac{\dot{\boldsymbol{\beta}}_t}{\boldsymbol{\beta}_t}\right) \odot \mathbf{x} \tag{60}$$

$$\epsilon_t(\mathbf{x} \mid \mathbf{z}) = \left(-\frac{\boldsymbol{\alpha}_t}{\boldsymbol{\beta}_t}\right) \odot \mathbf{z} + \left(-\frac{1}{\boldsymbol{\beta}_t}\right) \odot \mathbf{x} \tag{61}$$

$$z_t(\mathbf{x} \mid \mathbf{z}) = 1 \odot \mathbf{z} + 0 \odot \mathbf{x} \tag{62}$$

$$s_t(\mathbf{x} \mid \mathbf{z}) = \left(\frac{\boldsymbol{\alpha}_t}{\boldsymbol{\beta}_t^2}\right) \odot \mathbf{z} + \left(-\frac{1}{\boldsymbol{\beta}_t^2}\right) \odot \mathbf{x} \tag{63}$$

These prediction targets share a common form:

$$f_t(\mathbf{x} \mid \mathbf{z}) = \mathbf{a}_t \odot \mathbf{z} + \mathbf{b}_t \odot \mathbf{x} \tag{64}$$

where $\odot$ denotes element-wise multiplication. The corresponding marginal prediction conditioned on $\mathbf{c}$ is given by the posterior expectation:

$$f_t(\mathbf{x}, \mathbf{c}) = \int f_t(\mathbf{x} \mid \mathbf{z}) \frac{p_t(\mathbf{x} \mid \mathbf{z}) p_{\text{data}}(\mathbf{z} \mid \mathbf{c})}{p_t(\mathbf{x}, \mathbf{c})} \mathrm{d}\mathbf{z} \tag{65}$$

$$= \mathbf{a}_t \odot \int \mathbf{z} \frac{p_t(\mathbf{x} \mid \mathbf{z}) p_{\text{data}}(\mathbf{z} \mid \mathbf{c})}{p_t(\mathbf{x}, \mathbf{c})} \mathrm{d}\mathbf{z} + \mathbf{b}_t \odot \mathbf{x} \tag{66}$$

$$= \mathbf{a}_t \odot g_t(\mathbf{x}, \mathbf{c}) + \mathbf{b}_t \odot \mathbf{x} \tag{67}$$

Thus, the core task is to compute the posterior mean of the data $\mathbf{z}$, denoted as $g_t(\mathbf{x}, \mathbf{c}) \coloneqq \mathbb{E}[\mathbf{z} \mid \mathbf{x}, \mathbf{c}]$.

Under our specialized triangular schedule, we partition the sequence into three regions: finalized $(0 : m)$, active $(m : n)$, and future $(n : K)$. We denote the corresponding sub-vectors as $\mathbf{v}^1, \mathbf{v}^2, \mathbf{v}^3$ for any vector $\mathbf{v}$:

$$\mathbf{z} = [\mathbf{z}^{0:m}, \mathbf{z}^{m:n}, \mathbf{z}^{n:K}] = [\mathbf{z^1}, \mathbf{z^2}, \mathbf{z^3}] \tag{68}$$

$$\mathbf{x} = [\mathbf{x}^{0:m}, \mathbf{x}^{m:n}, \mathbf{x}^{n:K}] = [\mathbf{x^1}, \mathbf{x^2}, \mathbf{x^3}] \tag{69}$$

$$\mathbf{c} = [\mathbf{c}^{0:m}, \mathbf{c}^{m:n}, \mathbf{c}^{n:K}] = [\mathbf{c^1}, \mathbf{c^2}, \mathbf{c^3}] \tag{70}$$

The conditional distribution $p_t(\mathbf{x} \mid \mathbf{z})$ factorizes as:

$$p_t(\mathbf{x} \mid \mathbf{z}) = p_t(\mathbf{x^1} \mid \mathbf{z^1}) p_t(\mathbf{x^2} \mid \mathbf{z^2}) p_t(\mathbf{x^3} \mid \mathbf{z^3}) \tag{71}$$

$$= \delta(\mathbf{x^1} - \mathbf{z^1}) p_t(\mathbf{x^2} \mid \mathbf{z^2}) p_{\text{noise}}(\mathbf{x^3}) \tag{72}$$

The data prior $p_{\text{data}}(\mathbf{z} \mid \mathbf{c})$ factorizes according to the causal dependency (Assumption 3.7):

$$p_{\text{data}}(\mathbf{z} \mid \mathbf{c}) = p_{\text{data}}(\mathbf{z^1} \mid \mathbf{c^1}) p_{\text{data}}(\mathbf{z^2} \mid \mathbf{z^1}, \mathbf{c^1}, \mathbf{c^2}) \tag{73}$$

$$\times\, p_{\text{data}}(\mathbf{z^3} \mid \mathbf{z^1}, \mathbf{z^2}, \mathbf{c^1}, \mathbf{c^2}, \mathbf{c^3}) \tag{74}$$

We now compute the posterior mean $g_t(\mathbf{x}, \mathbf{c})$ for each region.

For Region 1 $(0 : m)$, since $p_t(\mathbf{x^1} \mid \mathbf{z^1})$ is a Dirac delta, we have:

$$g_t^{\mathbf{1}}(\mathbf{x}, \mathbf{c}) = \mathbf{x^1}. \tag{75}$$

For Region 2 $(m : n)$, the posterior depends on the history $\mathbf{z}^1 = \mathbf{x}^1$:

$$g_t^{\mathbf{2}}(\mathbf{x}, \mathbf{c}) = \int \mathbf{z^2} \frac{p_t(\mathbf{x^2} \mid \mathbf{z^2}) p_{\text{data}}(\mathbf{z^2} \mid \mathbf{z^1} = \mathbf{x^1}, \mathbf{c^{1,2}})}{p_t(\mathbf{x^2} \mid \mathbf{z^1} = \mathbf{x^1}, \mathbf{c^{1,2}})} \mathrm{d}\mathbf{z^2} \tag{76}$$

This term depends only on $\mathbf{x}^{1,2}$ and $\mathbf{c}^{1,2}$.

For Region 3 $(n : K)$, the posterior mean involves an expectation over $\mathbf{z}^3$. Note that $p_t(\mathbf{x}^3 \mid \mathbf{z}^3) = p_{\text{noise}}(\mathbf{x}^3)$ is independent of $\mathbf{z}^3$. Thus:

$$g_t^{\mathbf{3}}(\mathbf{x}, \mathbf{c}) = \int \mathbf{z^3} \frac{p_t(\mathbf{x} \mid \mathbf{z}) p_{\text{data}}(\mathbf{z} \mid \mathbf{c})}{p_t(\mathbf{x}, \mathbf{c})} \mathrm{d}\mathbf{z} \tag{77}$$

$$= \int \left[\int \mathbf{z^3} p_{\text{data}}(\mathbf{z^3} \mid \mathbf{z^1} = \mathbf{x^1}, \mathbf{z^2}, \mathbf{c}) \mathrm{d}\mathbf{z^3} \right] \tag{78}$$

$$\times \frac{p_t(\mathbf{x^2} \mid \mathbf{z^2}) p_{\text{data}}(\mathbf{z^2} \mid \mathbf{z^1} = \mathbf{x^1}, \mathbf{c^{1,2}})}{p_t(\mathbf{x^2} \mid \mathbf{x^1}, \mathbf{c^{1,2}})} \mathrm{d}\mathbf{z^2} \tag{79}$$

$$= \int \mathbb{E}[\mathbf{z^3} \mid \mathbf{z^1} = \mathbf{x^1}, \mathbf{z^2}, \mathbf{c}] \tag{80}$$

$$\times \frac{p_t(\mathbf{x^2} \mid \mathbf{z^2}) p_{\text{data}}(\mathbf{z^2} \mid \mathbf{z^1} = \mathbf{x^1}, \mathbf{c^{1,2}})}{p_t(\mathbf{x^2} \mid \mathbf{x^1}, \mathbf{c^{1,2}})} \mathrm{d}\mathbf{z^2} \tag{81}$$

However, in this region, $\alpha_t^k = 0$ and $\beta_t^k = 1$, implying $\dot{\boldsymbol{\alpha}}_t = \mathbf{0}$ and $\dot{\boldsymbol{\beta}}_t = \mathbf{0}$. Consequently, the velocity coefficients $\mathbf{a}_t$ and $\mathbf{b}_t$ are zero, so the value of g_t^3 does not affect the velocity field.

Finally, we substitute these results into the velocity equation. In Regions 1 and 3, the time derivatives vanish, yielding zero velocity. In Region 2, the velocity is determined by the local posterior mean. Thus:

$$u_t(\mathbf{X}_t, \mathbf{c}^{0:K}) \tag{82}$$

$$= \begin{bmatrix} \mathbf{0^1} \\ \left(\dot{\boldsymbol{\alpha}}_t - \frac{\dot{\boldsymbol{\beta}}_t}{\boldsymbol{\beta}_t} \odot \boldsymbol{\alpha}_t\right) \odot g_t^{\mathbf{2}}(\mathbf{X}_t^{\mathbf{1,2}}, \mathbf{c^{1,2}}) + \left(\frac{\dot{\boldsymbol{\beta}}_t}{\boldsymbol{\beta}_t}\right) \odot \mathbf{X}_t \\ \mathbf{0^3} \end{bmatrix} \tag{83}$$

$$= \begin{bmatrix} \mathbf{0}^{0:m(t)} \\ u_t^{m(t):n(t)}(\mathbf{X}_t^{0:n(t)}, \mathbf{c}^{0:n(t)}) \\ \mathbf{0}^{n(t):K} \end{bmatrix} \tag{84}$$

Regardless of the prediction parameterization, since the update $\mathrm{d}\mathbf{X}_t$ always reverts to the velocity u_t (which is zero in regions 1 and 3), we are only concerned with the active window. Thus, it suffices to train the prediction target within the active window.

□

Remark B.5 (Stochastic Extension)**.** *If we introduce a diffusion term $u_t^{fp} = u_t + \frac{\boldsymbol{\sigma}_t^2}{2} \odot s_t$, the drift u_t^{fp} is no longer guaranteed to be zero in the finalized and future regions. However, the score function s_t in these regions depends only on $\mathbf{X}_t$:*

- *In the future region ($k \geq n(t)$), we have $\alpha_t^k = 0, \beta_t^k = 1$. The posterior mean vanishes, so substituting into the score definition yields $s_t(\mathbf{x}) = -\mathbf{x}$.*
- *In the finalized region ($k < m(t)$), we substitute the posterior mean $g_t(\mathbf{x}) = \mathbf{x}$ derived in the proof. Keeping α_t and β_t (treating β_t as a small non-zero value to avoid*

	layers	hidden	ffn	heads	window	$v/\epsilon/x_0$	steps	FID↓	R@3↑	MM-Dist↓	Diversity→
real motion							-	0.002	0.797	2.974	9.503
ours	8	1024	2048	8	5	v	10	**0.057**$^{\pm.002}$	0.810$^{\pm.003}$	2.887$^{\pm.007}$	9.579$^{\pm.062}$
	2	1024	2048	8	5	v	10	0.087	0.787	3.006	9.419
	4	1024	2048	8	5	v	10	0.080	0.810	2.925	9.536
	12	1024	2048	8	5	v	10	0.062	0.798	2.942	**9.524**
	8	128	512	8	5	v	10	0.121	0.786	3.048	9.551
	8	256	1024	8	5	v	10	0.077	0.800	2.954	9.531
	8	512	2048	8	5	v	10	0.088	0.811	2.949	9.560
	8	768	3072	16	5	v	10	0.084	0.812	2.936	9.584
	8	1024	4096	16	5	v	10	0.073	0.814	2.909	9.504
	8	2048	8192	16	5	v	10	0.081	0.792	3.001	9.561
	8	1024	2048	8	1	v	10	1.38	0.644	3.987	8.855
	8	1024	2048	8	20	v	10	0.060	**0.821**	**2.842**	9.683
	8	1024	2048	8	5	ϵ	10	0.124	0.807	2.855	9.549
	8	1024	2048	8	5	x_0	10	0.060	0.805	2.901	9.525
	8	1024	2048	8	5	v	20	0.061	0.807	2.903	9.624
	8	1024	2048	8	5	v	100	0.060	0.812	2.895	9.653

Table 5. Ablation study of the model architecture on HumanML3D test set.

division by zero), the marginal score becomes

$$s_t(\mathbf{x}) = \frac{\boldsymbol{\alpha}_t \odot \mathbf{x} - \mathbf{x}}{\boldsymbol{\beta}_t^2} \tag{85}$$

$$= -\frac{(1-\boldsymbol{\alpha}_t) \odot \mathbf{x}}{\boldsymbol{\beta}_t^2} = -\frac{\mathbf{x}}{\boldsymbol{\beta}_t} \tag{86}$$

Since we generally do not wish to add diffusion to parts that are already denoised or remain pure noise, and to avoid numerical instability where $\beta_t \approx 0$, we set $\boldsymbol{\sigma}_t$ to be a diagonal matrix that is non-zero only in the active window. This preserves the streaming locality. Within the active window, s_t can be derived from u_t.

Remark B.6 (Interpretation of Causal Dependency)**.** *This assumption (Assumption 3.7) does not imply that the model cannot plan or execute complex behaviors. The available control signal $\mathbf{c}^{0:l}$ can itself contain complex, long-term instructions (e.g., "first walk, then run"). The condition merely states that the motion generated up to frame l does not depend on future, unseen instructions that arrive later in the stream. Thus, this assumption holds in most practical streaming scenarios. We explicitly use this assumption to factorize $p_{data}(\mathbf{z} \mid \mathbf{c})$ in the proof of Theorem B.4.*

C. Baseline Implementation Details

To ensure a fair comparison, we modified the PRIMAL baseline as follows: (1) replaced the cosine schedule-based diffusion training with standard flow matching; (2) changed the transformer backbone from `nn.TransformerEncoderBlock` to the DiT block used in Wan; (3) removed the FK and velocity losses; (4) upgraded the text conditioning from discrete action tags (*e.g.*, "walk", "run") to natural language descriptions using a T5 encoder; and (5) adopted the same 263D motion representation instead of the original 267D format.

D. Hyper Parameter Search

We conduct a grid search over the main hyperparameters of our motion generation network, and report the results in Table 5. A hidden size of 1024 yields the best performance in our setting. Increasing the window size provides slight gains, but at the cost of higher response latency. Different prediction types do not show significant performance differences under our configuration.

E. Details of User Study

To evaluate the perceptual quality of the generated motions, we conducted a user study with 100 participants. We compared our *FloodDiffusion* against real motion ground truth (GT) and two streaming baselines: *PRIMAL* [38] and *MotionStreamer* [30].

Questionnaire Design We collected 100 questionnaires in total. Each questionnaire consists of three distinct questions, where each question compares a randomly sampled pair of videos generated by two different methods. The three questions correspond to the three evaluation metrics respectively:

1. **Preference**: Given a pair of videos, choose the one that appears more reasonable and plausible given the text prompt.
2. **Transition**: Given a pair of videos, choose the one that transitions more smoothly between different actions.
3. **Consistency**: Given a pair of videos, choose the one that better maintains a consistent motion style across different actions.

An example of the question interface is shown in Figure 7.

Comparison (A vs B)	Preference Win Rate (A : B)	Transition Win Rate (A : B)	Consistency Win Rate (A : B)
Ours vs. Baselines			
FloodDiffusion vs PRIMAL	63.2% (12 : 7)	62.5% (10 : 6)	55.6% (10 : 8)
FloodDiffusion vs MotionStreamer	56.3% (9 : 7)	54.5% (6 : 5)	50.0% (6 : 6)
Ours vs. Real Motion			
FloodDiffusion vs GT	50.0% (8 : 8)	47.1% (8 : 9)	42.1% (8 : 11)
Reference Comparisons			
PRIMAL vs GT	31.8% (7 : 15)	33.3% (5 : 10)	42.9% (6 : 8)
MotionStreamer vs GT	28.6% (4 : 10)	40.0% (8 : 12)	39.1% (9 : 14)
MotionStreamer vs PRIMAL	61.5% (8 : 5)	52.4% (11 : 10)	57.1% (8 : 6)

Table 6. **Pairwise comparison results.** We report the win rate of method A over method B, along with the raw vote counts in parentheses. Our method outperforms both baselines on all metrics (or ties) and achieves a win rate close to 50% against real motion (GT), indicating high realism.

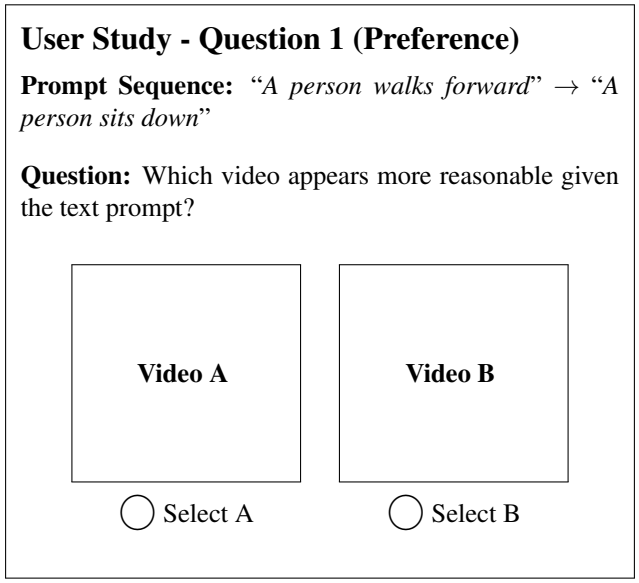

Figure 7. An illustrative example of a single question in the user study. Each questionnaire contains three such comparisons, one for each metric (Preference, Transition, and Consistency), using different video pairs.

Pairwise Comparison Results We aggregated the votes from all valid questionnaires. The detailed head-to-head win counts for each pair of methods are reported in Table 6. These raw counts were used to compute the Bradley–Terry scores presented in the main text.