

Diversity in flow-matching, bidirectional versus causal

Zhifan Zhu

October 15 2026

A bidirectional video generator, when implemented using flow matching and trained using MSE loss until optimum, has more diverse predicted velocities (measured in variance) than its causal version, e.g. those interactive, streaming world models. Here *bidirectional* and *causal* differ only in the access to the future context – bidirectional sees the future context while causal doesn't.

It is plausible that the final endpoints – the videos – are also more diverse in the bidirectional generations. But there seems to be no good mathematical justification for this final-endpoint diversity.

The velocity is more diverse because:

1. MSE optimum

The MSE objective is

$$L(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{c}, \mathbf{x}) \sim p_{\text{data}}(\mathbf{c}, \mathbf{x})} [\|f(\mathbf{c}; \boldsymbol{\theta}) - \mathbf{x}\|^2], \quad (1)$$

where $\boldsymbol{\theta}$ denotes the network parameters, $\mathbf{c}$ is context, $\mathbf{x}$ is ground-truth velocity, and $(\mathbf{c}, \mathbf{x})$ is a paired sample. For brevity, expectation subscripts below abbreviate the corresponding marginal or conditional distributions of p_{data} .

At the MSE optimum,

$$f(\mathbf{c}; \boldsymbol{\theta}^*) = \mathbb{E}_{\mathbf{x}|\mathbf{c}}[\mathbf{x}] = \mathbb{E}[\mathbf{x} | \mathbf{c}]. \quad (2)$$

The derivation:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{c}, \mathbf{x}} [\|f(\mathbf{c}; \boldsymbol{\theta}) - \mathbb{E}[\mathbf{x} | \mathbf{c}] + \mathbb{E}[\mathbf{x} | \mathbf{c}] - \mathbf{x}\|^2] \\ &= \mathbb{E}_{\mathbf{c}, \mathbf{x}} [\|f(\mathbf{c}; \boldsymbol{\theta}) - \mathbb{E}[\mathbf{x} | \mathbf{c}]\|^2] \\ &\quad + 2\mathbb{E}_{\mathbf{c}, \mathbf{x}} [\langle f(\mathbf{c}; \boldsymbol{\theta}) - \mathbb{E}[\mathbf{x} | \mathbf{c}], \mathbb{E}[\mathbf{x} | \mathbf{c}] - \mathbf{x} \rangle] \\ &\quad + \mathbb{E}_{\mathbf{c}, \mathbf{x}} [\|\mathbb{E}[\mathbf{x} | \mathbf{c}] - \mathbf{x}\|^2]. \end{aligned} \quad (3)$$

The third term is irrelevant to $\boldsymbol{\theta}$. And the second term is 0 because

$$\begin{aligned} &\mathbb{E}_{\mathbf{c}, \mathbf{x}} [\langle f(\mathbf{c}; \boldsymbol{\theta}) - \mathbb{E}[\mathbf{x} | \mathbf{c}], \mathbb{E}[\mathbf{x} | \mathbf{c}] - \mathbf{x} \rangle] \\ &= \mathbb{E}_{\mathbf{c}} \left[\mathbb{E}_{\mathbf{x}|\mathbf{c}} [\langle f(\mathbf{c}; \boldsymbol{\theta}) - \mathbb{E}[\mathbf{x} | \mathbf{c}], \mathbb{E}[\mathbf{x} | \mathbf{c}] - \mathbf{x} \rangle] \right] \\ &= \mathbb{E}_{\mathbf{c}} \left[\langle f(\mathbf{c}; \boldsymbol{\theta}) - \mathbb{E}[\mathbf{x} | \mathbf{c}], \mathbb{E}_{\mathbf{x}|\mathbf{c}} [\mathbb{E}[\mathbf{x} | \mathbf{c}] - \mathbf{x}] \rangle \right] \\ &= \mathbb{E}_{\mathbf{c}} [\langle f(\mathbf{c}; \boldsymbol{\theta}) - \mathbb{E}[\mathbf{x} | \mathbf{c}], \mathbb{E}[\mathbf{x} | \mathbf{c}] - \mathbb{E}[\mathbf{x} | \mathbf{c}] \rangle] \\ &= 0. \end{aligned} \quad (4)$$

Note that $\mathbb{E}[\mathbf{x} | \mathbf{c}]$ is a function of $\mathbf{c}$ only, not $\mathbf{x}$.

2. Bidirectional and causal contexts

For the bidirectional generator, the context $\mathbf{c}$ is $(\mathbf{u}, \mathbf{f})$, where $\mathbf{u}$ is past context and $\mathbf{f}$ is future context. For the causal generator, the context $\mathbf{c}$ is $\mathbf{u}$ only. At the MSE optimum,

$$f_{\text{bi}}(\mathbf{u}, \mathbf{f}; \boldsymbol{\theta}^*) = E_{\mathbf{x}|\mathbf{u}, \mathbf{f}}[\mathbf{x}] = E[\mathbf{x} | \mathbf{u}, \mathbf{f}], \quad (5)$$

whereas

$$f_{\text{caus}}(\mathbf{u}; \boldsymbol{\theta}^*) = E_{\mathbf{x}|\mathbf{u}}[\mathbf{x}] = E[\mathbf{x} | \mathbf{u}]. \quad (6)$$

These two terms are related by $E_{\mathbf{f}|\mathbf{u}}[E[\mathbf{x} | \mathbf{u}, \mathbf{f}] | \mathbf{u}] = E[\mathbf{x} | \mathbf{u}]$ (by the law of iterated expectation). This means that $f_{\text{caus}}(\mathbf{u}; \boldsymbol{\theta}^*)$ is the conditional expectation of $f_{\text{bi}}(\mathbf{u}, \mathbf{f}; \boldsymbol{\theta}^*)$:

$$f_{\text{caus}}(\mathbf{u}; \boldsymbol{\theta}^*) = E_{\mathbf{f}|\mathbf{u}}[f_{\text{bi}}(\mathbf{u}, \mathbf{f}; \boldsymbol{\theta}^*)]. \quad (7)$$

The variance of the bidirectionally predicted velocity is

$$\text{Var}(f_{\text{bi}}(\mathbf{u}, \mathbf{f}; \boldsymbol{\theta}^*)), \quad (8)$$

and the variance of the causally predicted velocity is

$$\text{Var}(f_{\text{caus}}(\mathbf{u}; \boldsymbol{\theta}^*)), \quad (9)$$

which is equivalent to $\text{Var}(E[f_{\text{bi}}(\mathbf{u}, \mathbf{f}; \boldsymbol{\theta}^*) | \mathbf{u}])$.

3. The law of total variance

The law of total variance says $\text{Var}(Y) = \text{Var}(E[Y | X]) + E[\text{Var}(Y | X)]$. In other words, $\text{Var}(Y) \geq \text{Var}(E[Y | X])$.

Replace Y with $f_{\text{bi}}(\mathbf{u}, \mathbf{f}; \boldsymbol{\theta}^*)$ and X with $\mathbf{u}$. Then

$$\begin{aligned} \text{Var}(f_{\text{bi}}(\mathbf{u}, \mathbf{f}; \boldsymbol{\theta}^*)) &\geq \text{Var}(E[f_{\text{bi}}(\mathbf{u}, \mathbf{f}; \boldsymbol{\theta}^*) | \mathbf{u}]) \\ &= \text{Var}(f_{\text{caus}}(\mathbf{u}; \boldsymbol{\theta}^*)). \end{aligned} \quad (10)$$

Note that the inequality allows equality, so rigorously we should say that bidirectional is at least as diverse as causal.

Acknowledgement. I suspected that bidirectional produces more diversity, and I asked ChatGPT. I am mathematically convinced.