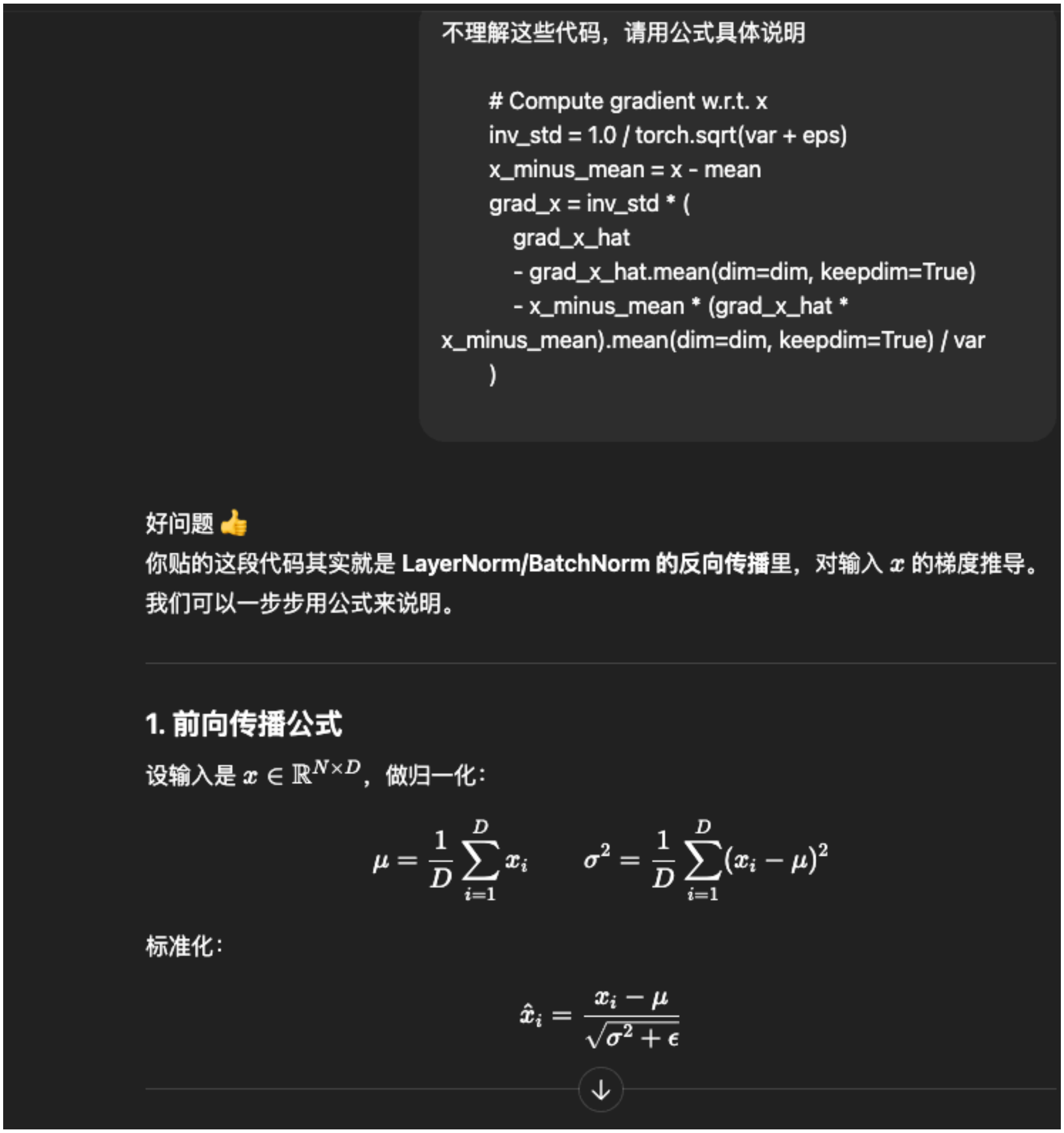


不理解这些代码，请用公式具体说明

```
    # Compute gradient w.r.t. x
    inv_std = 1.0 / torch.sqrt(var + eps)
    x_minus_mean = x - mean
    grad_x = inv_std * (
      grad_x_hat
      - grad_x_hat.mean(dim=dim, keepdim=True)
      - x_minus_mean * (grad_x_hat * 
x_minus_mean).mean(dim=dim, keepdim=True) / var
    )
```

好问题 👍

你贴的这段代码其实就是 **LayerNorm/BatchNorm 的反向传播里**，对输入 x 的梯度推导。我们可以一步步用公式来说明。

---

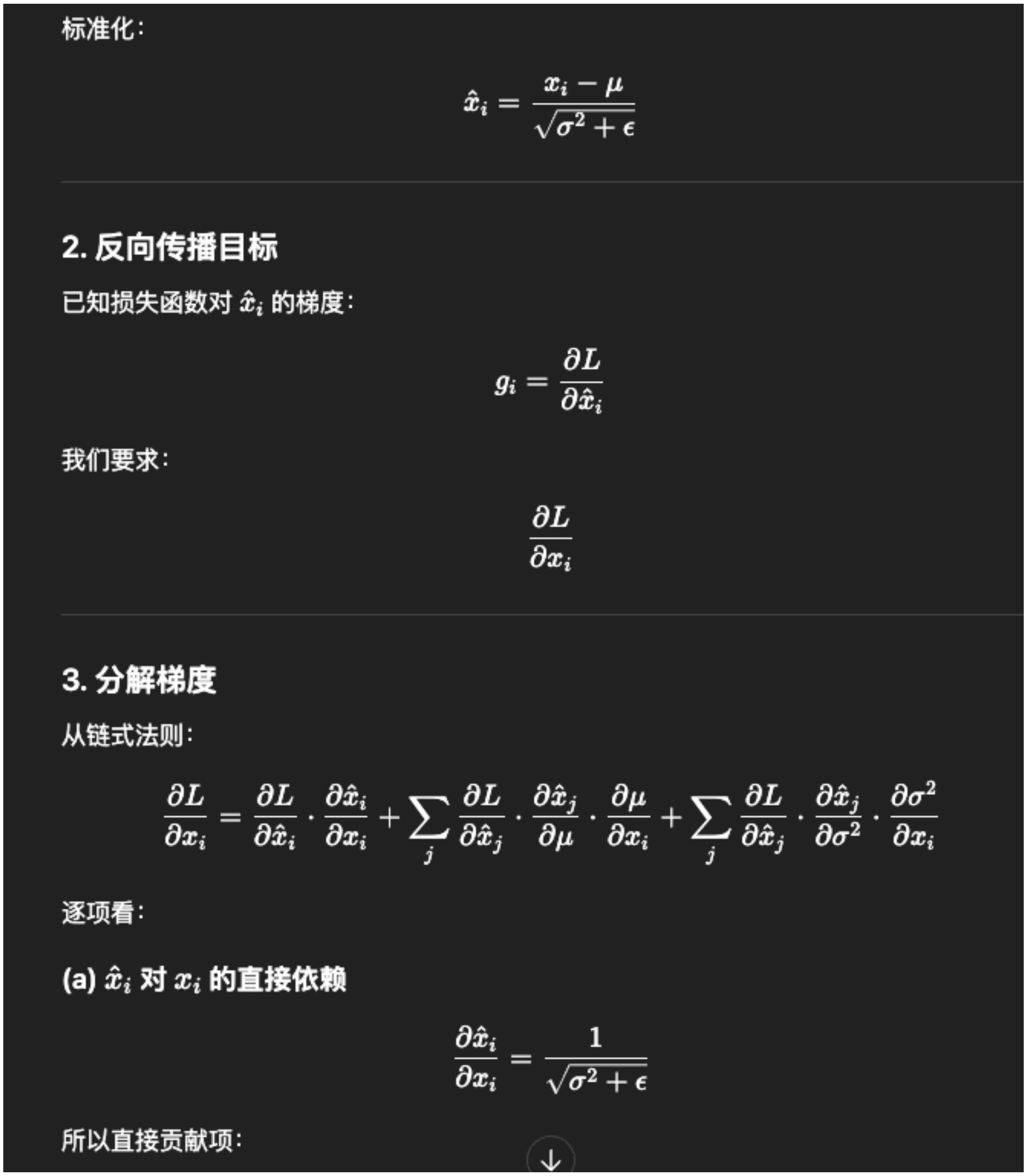
1. 前向传播公式

设输入是 $x \in \mathbb{R}^{N \times D}$，做归一化：

$$\mu = \frac{1}{D}\sum_{i=1}^{D} x_i \qquad \sigma^2 = \frac{1}{D}\sum_{i=1}^{D}(x_i - \mu)^2$$

标准化：

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$



标准化：

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

---

2. 反向传播目标

已知损失函数对 $\hat{x}_i$ 的梯度：

$$g_i = \frac{\partial L}{\partial \hat{x}_i}$$

我们要求：

$$\frac{\partial L}{\partial x_i}$$

---

3. 分解梯度

从链式法则：

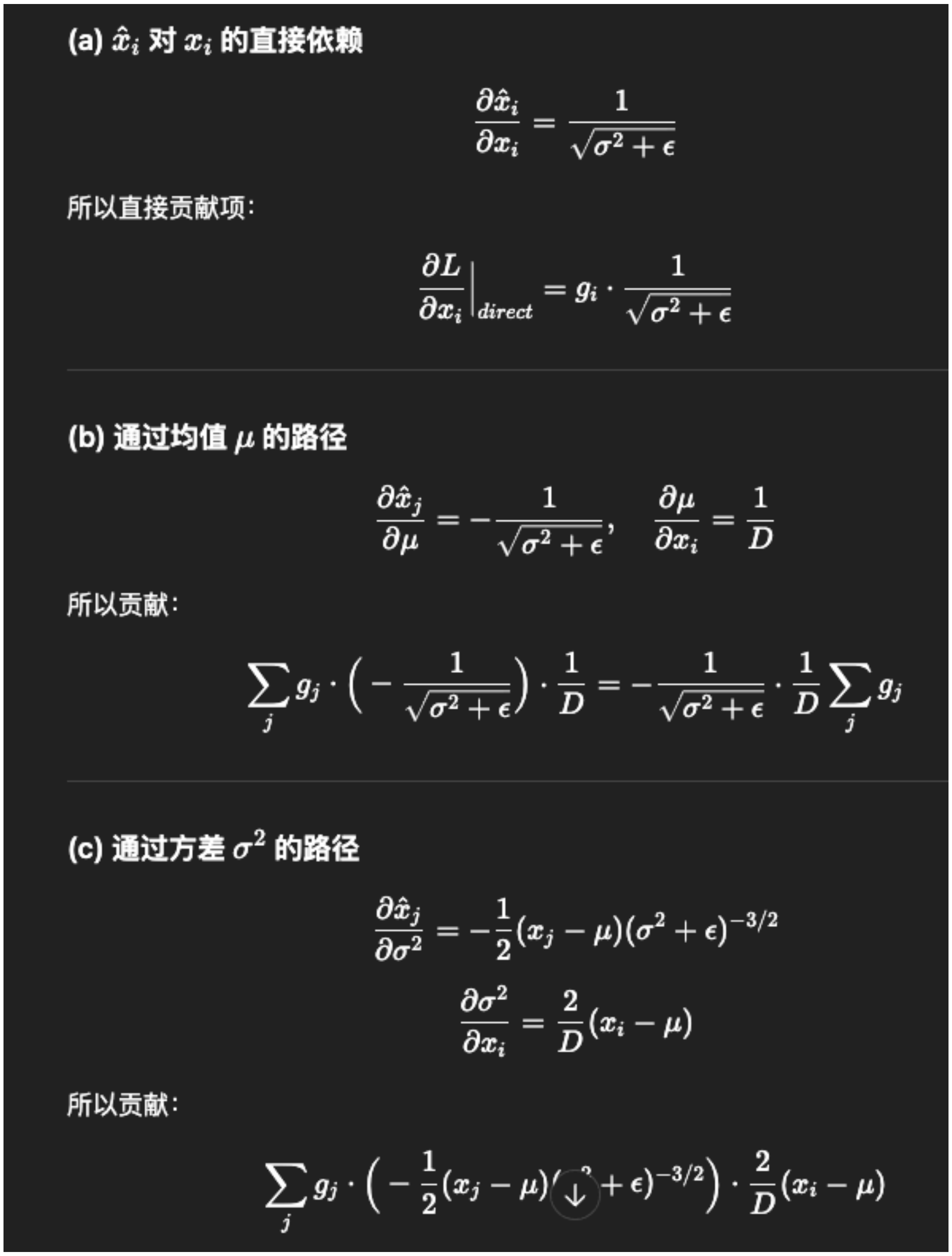
$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{\partial \hat{x}_i}{\partial x_i} + \sum_j \frac{\partial L}{\partial \hat{x}_j} \cdot \frac{\partial \hat{x}_j}{\partial \mu} \cdot \frac{\partial \mu}{\partial x_i} + \sum_j \frac{\partial L}{\partial \hat{x}_j} \cdot \frac{\partial \hat{x}_j}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial x_i}$$

逐项看：

(a) $\hat{x}_i$ 对 x_i 的直接依赖

$$\frac{\partial \hat{x}_i}{\partial x_i} = \frac{1}{\sqrt{\sigma^2 + \epsilon}}$$

所以直接贡献项：



(a) $\hat{x}_i$ 对 x_i 的直接依赖

$$\frac{\partial \hat{x}_i}{\partial x_i} = \frac{1}{\sqrt{\sigma^2 + \epsilon}}$$

所以直接贡献项：

$$\frac{\partial L}{\partial x_i}\Big|_{direct} = g_i \cdot \frac{1}{\sqrt{\sigma^2 + \epsilon}}$$

(b) 通过均值 μ 的路径

$$\frac{\partial \hat{x}_j}{\partial \mu} = -\frac{1}{\sqrt{\sigma^2 + \epsilon}}, \quad \frac{\partial \mu}{\partial x_i} = \frac{1}{D}$$

所以贡献：

$$\sum_j g_j \cdot \left(-\frac{1}{\sqrt{\sigma^2 + \epsilon}}\right) \cdot \frac{1}{D} = -\frac{1}{\sqrt{\sigma^2 + \epsilon}} \cdot \frac{1}{D} \sum_j g_j$$

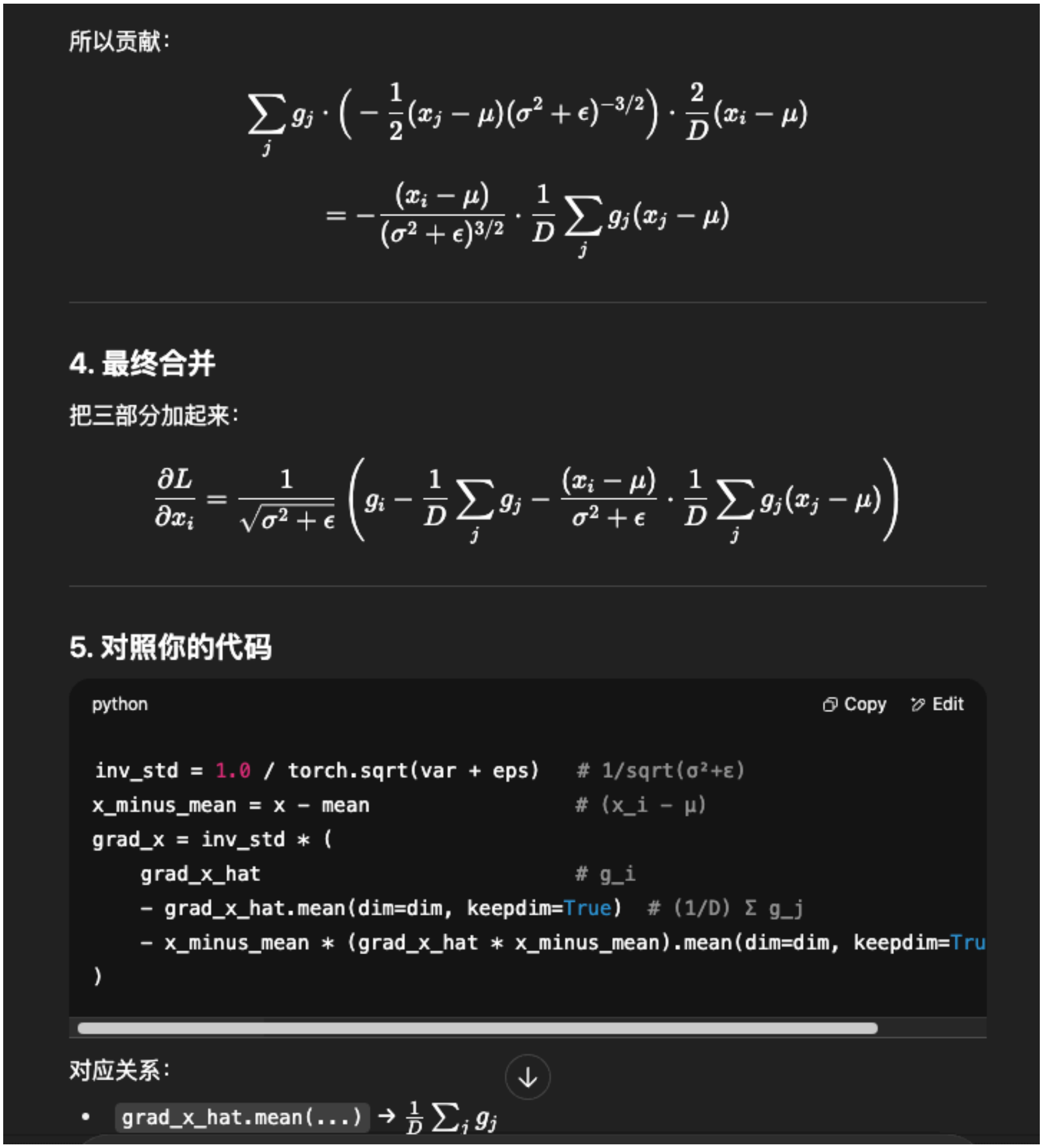
(c) 通过方差 σ^2 的路径

$$\frac{\partial \hat{x}_j}{\partial \sigma^2} = -\frac{1}{2}(x_j - \mu)(\sigma^2 + \epsilon)^{-3/2}$$

$$\frac{\partial \sigma^2}{\partial x_i} = \frac{2}{D}(x_i - \mu)$$

所以贡献：

$$\sum_j g_j \cdot \left(-\frac{1}{2}(x_j - \mu)(\sigma^2 + \epsilon)^{-3/2}\right) \cdot \frac{2}{D}(x_i - \mu)$$



所以贡献：

$$\sum_j g_j \cdot \left(-\frac{1}{2}(x_j-\mu)(\sigma^2+\epsilon)^{-3/2}\right)\cdot\frac{2}{D}(x_i-\mu)$$

$$= -\frac{(x_i-\mu)}{(\sigma^2+\epsilon)^{3/2}}\cdot\frac{1}{D}\sum_j g_j(x_j-\mu)$$

---

4. 最终合并

把三部分加起来：

$$\frac{\partial L}{\partial x_i} = \frac{1}{\sqrt{\sigma^2+\epsilon}}\left(g_i - \frac{1}{D}\sum_j g_j - \frac{(x_i-\mu)}{\sigma^2+\epsilon}\cdot\frac{1}{D}\sum_j g_j(x_j-\mu)\right)$$

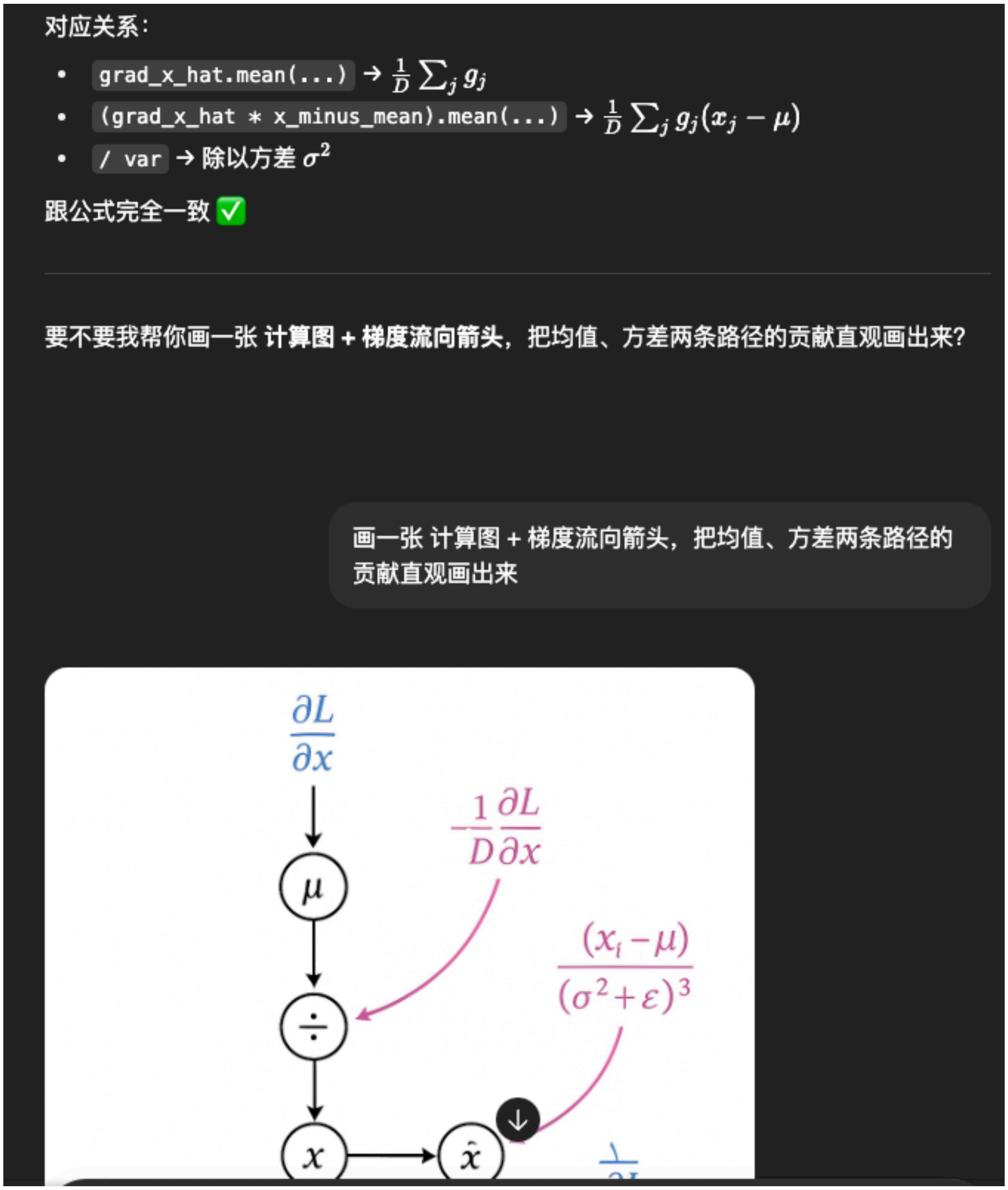
---

5. 对照你的代码

```python
inv_std = 1.0 / torch.sqrt(var + eps)   # 1/sqrt(σ²+ε)
x_minus_mean = x - mean                 # (x_i - μ)
grad_x = inv_std * (
    grad_x_hat                          # g_i
    - grad_x_hat.mean(dim=dim, keepdim=True)  # (1/D) Σ g_j
    - x_minus_mean * (grad_x_hat * x_minus_mean).mean(dim=dim, keepdim=Tru
)
```

对应关系：

- `grad_x_hat.mean(...)` → $\frac{1}{D}\sum_j g_j$



对应关系：

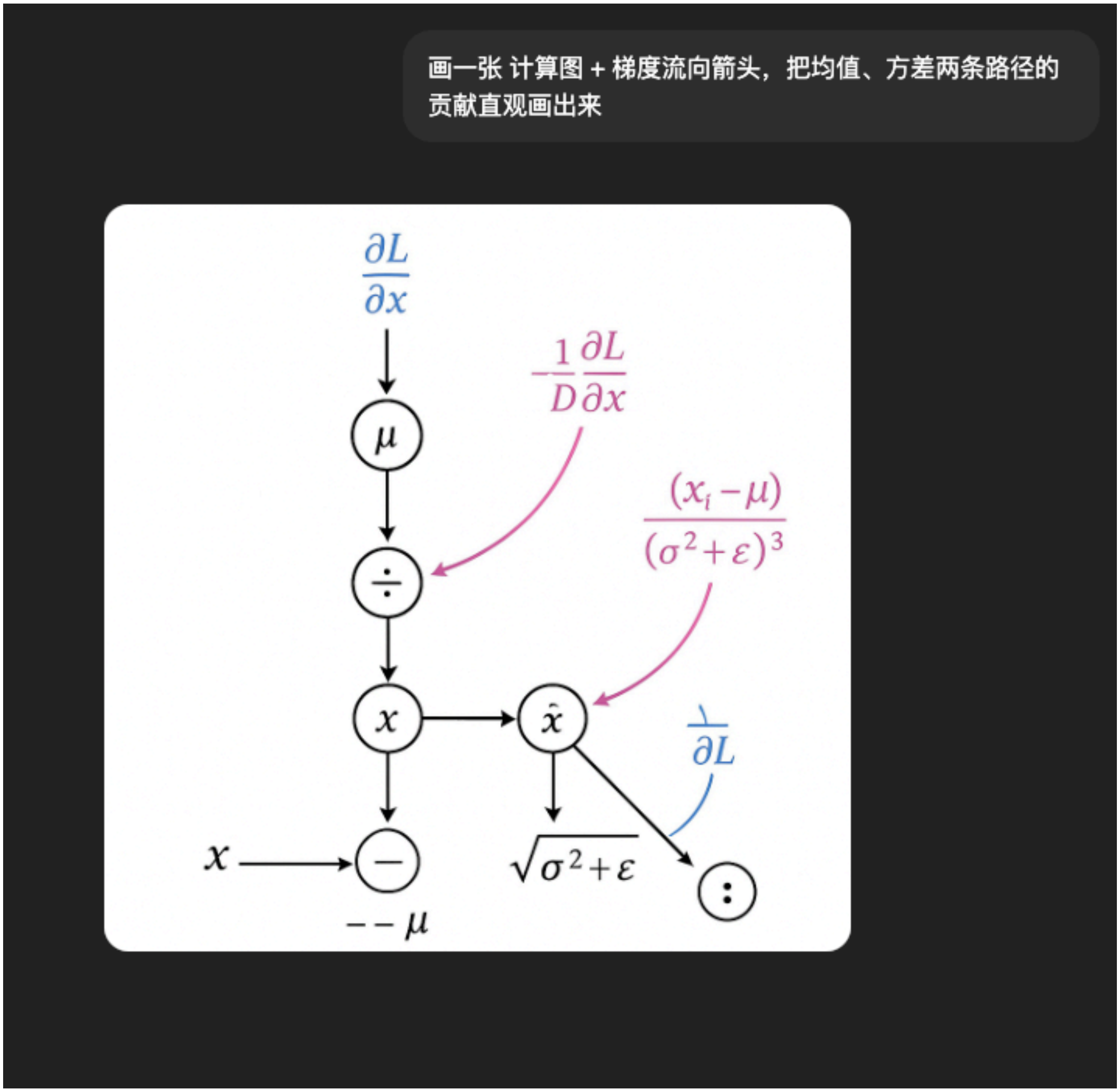
- `grad_x_hat.mean(...)` → $\frac{1}{D}\sum_j g_j$
- `(grad_x_hat * x_minus_mean).mean(...)` → $\frac{1}{D}\sum_j g_j(x_j - \mu)$
- `/ var` → 除以方差 σ^2

跟公式完全一致 ✅

要不要我帮你画一张 **计算图 + 梯度流向箭头**，把均值、方差两条路径的贡献直观画出来？

画一张 计算图 + 梯度流向箭头，把均值、方差两条路径的贡献直观画出来

$\frac{\partial L}{\partial x}$
$-\frac{1}{D}\frac{\partial L}{\partial x}$
$\frac{(x_i-\mu)}{(\sigma^2+\varepsilon)^3}$
μ
$\div$
x
$\hat{x}$



画一张 计算图 + 梯度流向箭头，把均值、方差两条路径的贡献直观画出来

$\frac{\partial L}{\partial x}$
μ
$-\frac{1}{D}\frac{\partial L}{\partial x}$
$\div$
$\frac{(x_i-\mu)}{(\sigma^2+\varepsilon)^3}$
x
$\hat{x}$
∂L
x
$-$
$--\mu$
$\sqrt{\sigma^2+\varepsilon}$
$:$