10 算例 G:《多模态数据下的 VAE 变种——多模态变分自编码器 (MVAE) 和模型升级设计》

在标准 VAE（算例 C）中，我们处理的是单一模态的数据观测 X。然而，客观世界的观测往往是多视角的：例如，描述同一个物体的“图像”和“文字”。多模态变分自编码器（MVAE, Wu & Goodman, 2018）的核心野心是：寻找一个共享的隐空间 Z，使得这个 Z 既能生成模态 1，也能生成模态 2，并且能够在模态缺失时实现跨模态推理。

10.1 MVAE 的概率图模型：生成假设与条件独立性

假设我们拥有两个模态的数据观测 X_1 和 X_2（例如 X_1 为图像，X_2 为对应文本），它们由同一个底层的隐变量 Z 驱动生成。

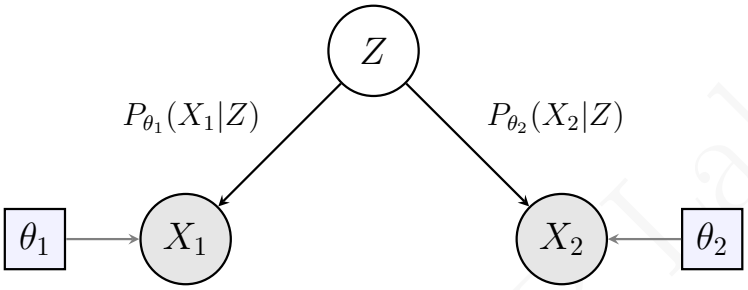

图 12: MVAE 的生成概率图模型 (PGM)。隐变量 Z 充当“公共原因”(Common Cause) 的角色。方框节点 θ_1 和 θ_2 分别代表两个独立解码器的网络参数。一旦 Z 被观测或给定，信息流就会被阻断，从而保证了 X_1 和 X_2 的条件独立性。

在生成模型（Generative Model）的设定中，MVAE 提出了一个至关重要的条件独立性假设 (Conditional Independence Assumption)：一旦给定了隐变量 Z，观测值 X_1 和 X_2 的生成过程是相互独立的。这意味着数据层面的所有相关性，都已经被 Z 这个枢纽完美地吸收了。

因此，联合生成概率可以分解为：

$$P_\theta(X_1, X_2, Z) = P(Z)P_{\theta_1}(X_1|Z)P_{\theta_2}(X_2|Z) \tag{10-1}$$

其中：

- $P(Z)$ 是共享先验，通常设为标准正态分布 $\mathcal{N}(0, \mathbf{I})$。
- P_{θ_1} 和 P_{θ_2} 分别是两个模态的生成网络（解码器）。

10.2 专家乘积 (Product of Experts, PoE) 联合推断

为了训练这个生成模型，我们需要最大化对数边缘似然 $\log P(X_1, X_2)$ 的变分下界（ELBO）。这要求我们引入一个联合推断网络 $q_\phi(Z|X_1, X_2)$ 来近似真实的后验 $P(Z|X_1, X_2)$。那么，如何合理地提出有关该推断网络的构造呢？MVAE 在这里提出一个有别于 Meanfield 的假设——它想法利用了贝叶斯公式的一些观察性结论：

$$P(Z|X_1, X_2) = \frac{P(X_1, X_2|Z)P(Z)}{P(X_1, X_2)} = \frac{P(X_1|Z)P(X_2|Z)P(Z)}{P(X_1, X_2)} \tag{10-2}$$

进一步，我们将单模态的似然 $P(X_i|Z)$ 替换为与单模态后验 $P(Z|X_i)$ 成比例的表达式：$P(X_i|Z) = \frac{P(Z|X_i)P(X_i)}{P(Z)}$，代入上式后化简可得：

$$P(Z|X_1, X_2) \propto \frac{P(Z|X_1)}{P(Z)}\frac{P(Z|X_2)}{P(Z)}P(Z) \propto P(Z)^{-1}P(Z|X_1)P(Z|X_2) \tag{10-3}$$

在变分推断中，我们用由参数 ϕ 定义的神经网络来近似这些分布。这引出了 MVAE 著名的 PoE 联合推断公式：[17]

[17]这是一个仅凭直觉的做法，并不建立在严格的数学推理上。

专家乘积 (PoE) 聚合机制

$$q_\phi(Z|X_1,X_2) \propto p(Z)\prod_{i=1}^{2} q_{\phi_i}(Z|X_i)$$

其中，$q_{\phi_i}(Z|X_i)$ 是专门处理第 i 个模态的独立推断网络（即“**专家**”）。

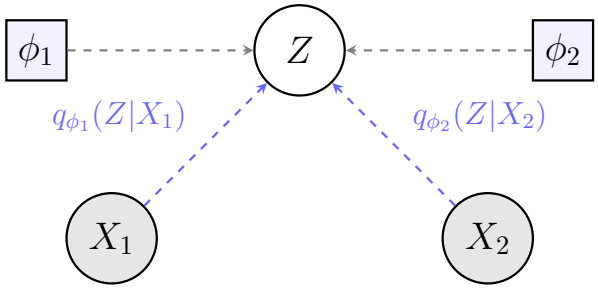

图 13: MVAE 的推断概率图模型 (PGM)。在此阶段，信息流从观测值反向流向隐空间，X_1 和 X_2 共同指向 Z，形成了一个“对撞节点”(Collider)。蓝色的虚线表示变分近似推断。独立的编码器网络（由 ϕ_1 和 ϕ_2 参数化）作为“专家”，在这里通过 PoE 机制将各自的推断结果进行汇流融合。

定理 10-1. *(高斯专家乘积 (PoE) 的闭式解析解)*

在多模态变分自编码器中，假设隐变量空间为 D 维。若先验 $p(\mathbf{Z})$ 与 N 个单模态推断网络 $q_{\phi_i}(\mathbf{Z}|\mathbf{X}_i)$ 均服从对角高斯分布，即协方差矩阵为对角阵 $diag(\boldsymbol{\sigma}^2)$。

那么，其联合推断分布（专家乘积）依然是一个对角高斯分布 $\mathcal{N}(\boldsymbol{\mu}_{joint}, diag(\boldsymbol{\sigma}^2_{joint}))$，其参数的闭式解为（下述运算均为向量的逐元素运算）：

$$\begin{cases} \dfrac{1}{\boldsymbol{\sigma}^2_{joint}} &= \dfrac{1}{\boldsymbol{\sigma}^2_{prior}} + \displaystyle\sum_{i=1}^{N} \dfrac{1}{\boldsymbol{\sigma}^2_i} \\ \boldsymbol{\mu}_{joint} &= \boldsymbol{\sigma}^2_{joint} \odot \left(\dfrac{\boldsymbol{\mu}_{prior}}{\boldsymbol{\sigma}^2_{prior}} + \displaystyle\sum_{i=1}^{N} \dfrac{\boldsymbol{\mu}_i}{\boldsymbol{\sigma}^2_i} \right) \end{cases} \tag{10-4}$$

意义解读：由于维度间相互独立，多维特征融合被解耦为 D 个独立的标量融合过程。联合精度为各专家精度之和，联合均值为各专家均值的精度加权平均。哪个专家越自信（方差越小），联合决策就越倾向于谁。

【定理证明】由于对角高斯分布的概率密度函数可分解为各维度边缘分布的连乘 $p(\mathbf{Z}) = \prod_{d=1}^{D} p(z_d)$，各维度间的推断是完全解耦的。因此，我们只需在单一维度的标量空间 z 下进行推导，所得结论可直接通过逐元素（element-wise）操作推广至全维向量空间。

设第 k 个分布（包含先验和所有专家）的概率密度函数为：

$$f_k(z) \propto \exp\left(-\frac{(z-\mu_k)^2}{2\sigma_k^2}\right)$$

其乘积分布 $q(z)$ 的指数部分为各分布指数项之和：

$$q(z) \propto \prod_{k=1}^{K} \exp\left(-\frac{(z-\mu_k)^2}{2\sigma_k^2}\right) = \exp\left(-\frac{1}{2}\sum_{k=1}^{K}\frac{(z-\mu_k)^2}{\sigma_k^2}\right)$$

将求和项中的二次型展开：

$$\begin{aligned} \sum_{k=1}^{K}\frac{(z-\mu_k)^2}{\sigma_k^2} &= \sum_{k=1}^{K}\left(\frac{1}{\sigma_k^2}z^2 - \frac{2\mu_k}{\sigma_k^2}z + \frac{\mu_k^2}{\sigma_k^2}\right) \\ &= \left(\sum_{k=1}^{K}\frac{1}{\sigma_k^2}\right)z^2 - 2\left(\sum_{k=1}^{K}\frac{\mu_k}{\sigma_k^2}\right)z + \sum_{k=1}^{K}\frac{\mu_k^2}{\sigma_k^2} \end{aligned}$$

对于目标高斯分布 $\mathcal{N}(\mu_{joint}, \sigma^2_{joint})$，其指数项的标准展开式为：

$$\frac{(z-\mu_{joint})^2}{\sigma^2_{joint}} = \frac{1}{\sigma^2_{joint}} z^2 - \frac{2\mu_{joint}}{\sigma^2_{joint}} z + \frac{\mu^2_{joint}}{\sigma^2_{joint}}$$

对比 z^2 与 z 的系数（系数匹配法）：

$$\begin{cases} \dfrac{1}{\sigma^2_{joint}} = \displaystyle\sum_{k=1}^{K} \frac{1}{\sigma^2_k} & \Longrightarrow \text{证明了精度相加性质} \\ \dfrac{\mu_{joint}}{\sigma^2_{joint}} = \displaystyle\sum_{k=1}^{K} \frac{\mu_k}{\sigma^2_k} & \Longrightarrow \mu_{joint} = \sigma^2_{joint} \left(\sum_{k=1}^{K} \frac{\mu_k}{\sigma^2_k} \right) \end{cases}$$

将上述标量结论应用至每个维度 $d \in \{1, \ldots, D\}$，并写成向量形式，即可得证。

10.3 MVAE 的 ELBO& 多模态场景下 ELBO 优化的策略调整

10.3.1 延用 VAE 的 ELBO 计算策略到多模态场景

有了联合推断网络 $q_\phi(Z|X_1, X_2)$，我们可以按照标准变分推导的习惯，从联合分布与近似后验的对数比值的期望出发，逐步展开双模态联合的变分下界 $\mathcal{L}(X_1, X_2)$。

根据 ELBO 的经典定义，我们将 MVAE 生成模型的条件独立性假设

$$P_\theta(X_1, X_2, Z) = P(Z) P_{\theta_1}(X_1|Z) P_{\theta_2}(X_2|Z)$$

代入分子中进行拆解，得到下式，请与 (eq.6-4) 进行比较：

$$\begin{aligned} \mathcal{L}(X_1, X_2) &= \mathbb{E}_{q_\phi(Z|X_1,X_2)} \left[\log \frac{P_\theta(X_1, X_2, Z)}{q_\phi(Z|X_1, X_2)} \right] \\ &= \mathbb{E}_{q_\phi(Z|X_1,X_2)} \left[\log \frac{P(Z) P_{\theta_1}(X_1|Z) P_{\theta_2}(X_2|Z)}{q_\phi(Z|X_1, X_2)} \right] \\ &= \mathbb{E}_{q_\phi(Z|X_1,X_2)} \left[\log P_{\theta_1}(X_1|Z) + \log P_{\theta_2}(X_2|Z) + \log \frac{P(Z)}{q_\phi(Z|X_1, X_2)} \right] \\ &= \underbrace{\mathbb{E}_{q_\phi(Z|X_1,X_2)} \left[\log P_{\theta_1}(X_1|Z) + \log P_{\theta_2}(X_2|Z) \right]}_{\text{双模态联合重建似然}} - \underbrace{\mathcal{D}_{KL}(q_\phi(Z|X_1, X_2) \parallel P(Z))}_{\text{联合后验与先验的 KL 散度}} \end{aligned} \tag{10-5}$$

通过这种展开，我们可以清晰地看到：MVAE 的联合下界本质上就是要求从融合后的隐变量 Z 中，同时以最大的概率重建出模态 X_1 和模态 X_2，并同时约束这个融合后的分布不要偏离先验 $P(Z)$。

与单模态 VAE 完全相同，公式中的期望项 $\mathbb{E}[\log P_\theta(X|Z)]$ 在代码中并不直接计算概率，而是转化为具体的重构损失函数。如果模态数据（如图像像素）被假设为高斯分布，最大化 $\log P_\theta(X|Z)$ 在数学上等价于最小化 **均方误差 (MSE)**：$\|X - \hat{X}\|_2^2$(见 eq.6-3)。。如果模态数据（如离散的文本 Token 或二值化特征）被假设为伯努利分布，则等价于最小化 **二元交叉熵 (BCE)**(见 eq.6-2)。

从逻辑而言，余下的工程实施与 VAE 类似，读者可以参照 VAE 相关章节。往下，我们重点关注多模态场景下的若干变化。

10.3.2 多模态场景下工程挑战之：模态间数据维度差异

在标准 VAE 中，Z 的采样仅受单一输入 X 的驱动。而在 MVAE 的联合训练阶段，隐变量 $Z \sim q_\phi(Z|X_1, X_2)$ 是通过前面推导的 **PoE（专家乘积）**融合而来的。这意味着，解码器提取梯度时，梯度会同时从 θ_1 和 θ_2 两个网络回传到同一个 Z 上。网络必须学会一种“妥协”：隐空间中的特征不仅要能完美还原图像，还要能准确拼凑出对应的文本。任何偏科（只讨好其中一个解码器）的行为都会被联合 ELBO 惩罚。

所以，MVAE 在工程实现上与 VAE 存在不同。在标准 VAE 中，重构误差与 KL 散度之间的平衡（如 β-VAE）已经足够棘手。但在 MVAE 中，不同模态的**维度大小往往存在天壤之别**。例如，X_1 是一张 256×256 的高维图像，X_2 只是一个 300 维的文本向量。如果不加干预直接相加，高维模态（图像）的 MSE 损失数值将呈压倒性优势，彻底淹没低维模态（文本）的梯度，导致模型退化为“仅仅在做图像重建，而无视文本”。因此，在实际代码中，联合下界通常会被改写为带有超参数平衡权重的形式：

$$\mathcal{L} \approx \lambda_1 \text{MSE}(X_1, \hat{X}_1) + \lambda_2 \text{BCE}(X_2, \hat{X}_2) + \beta \mathcal{D}_{KL}(q_\phi \parallel P(Z)) \tag{10-6}$$

通过精心调节 λ_1 和 λ_2，强制模型对不同维度的模态分配同等的注意力。

10.3.3 多模态场景下的工程挑战之：某模态数据缺失

工程挑战在于：在现实场景中，我们不仅需要同时拥有 X_1 和 X_2 时的联合推断能力，还需要在缺失某一模态时，仅凭 X_1 推断 Z（即 $q(Z|X_1)$），或仅凭 X_2 推断 Z（即 $q(Z|X_2)$）。如果硬训练一个同时接收 X_1 和 X_2 的庞大网络，当其中一个输入缺失时，网络将彻底瘫痪。

——简言之，如果只优化 $\mathcal{L}(X_1, X_2)$，模型只有在同时看到两个模态时才能正确工作，单模态的推断网络 q_{ϕ_1} 和 q_{ϕ_2} 并未得到充分的独立训练。

为此，MVAE 提出了子采样目标函数。我们不仅要求模型能进行联合推断，还强制要求每一个单独的模态也能完成对整体的推断与生成。最终的训练目标是最大化联合 ELBO 与所有边际 ELBO 的总和：

MVAE 的子采样策略

总体目标：最大化以下各项之和

$$\mathcal{J} = \mathcal{L}(X_1, X_2) + \mathcal{L}(X_1) + \mathcal{L}(X_2)$$

展开各项：

1. 联合 ELBO（双路输入，双路重建）：

$$\mathcal{L}(X_1, X_2) = \mathbb{E}_{q_\phi(Z|X_1,X_2)}[\log P_{\theta_1}(X_1|Z) + \log P_{\theta_2}(X_2|Z)] - \mathcal{D}_{KL}(q_\phi(Z|X_1, X_2) \parallel P(Z))$$

2. 边际 ELBO 1（仅用 X_1 推断，要求重建 X_1 和 X_2）：

$$\mathcal{L}(X_1) = \mathbb{E}_{q_{\phi_1}(Z|X_1)}[\log P_{\theta_1}(X_1|Z) + \log P_{\theta_2}(X_2|Z)] - \mathcal{D}_{KL}(q_{\phi_1}(Z|X_1) \parallel P(Z))$$

3. 边际 ELBO 2（仅用 X_2 推断，要求重建 X_1 和 X_2）：

$$\mathcal{L}(X_2) = \mathbb{E}_{q_{\phi_2}(Z|X_2)}[\log P_{\theta_1}(X_1|Z) + \log P_{\theta_2}(X_2|Z)] - \mathcal{D}_{KL}(q_{\phi_2}(Z|X_2) \parallel P(Z))$$

由于在训练中每次都精确计算所有组合的开销巨大，实际代码中通常采用模态 Dropout ：在每一个 Mini-batch 中，以一定的概率随机丢弃 X_1 或 X_2，迫使模型在动态缺损的条件下，依然拼尽全力通过 PoE 整合剩余信息，并对齐隐空间。

10.4 模型评估

10.4.1 对数重要性权重的方差 (Variance of the log importance weights)

> “In all VAE-family models, the inference network functions as an importance distribution for approximating the intractable posterior. A better importance distribution... results in importance weights with lower variance.” ——*MVAE, Wu & Goodman, 2018*

在重要性采样中，对于给定的数据观测 X 和采样得到的隐变量 $Z \sim q_\phi(Z|X)$，我们定义重要性权重 $w(Z)$ 为联合分布与近似后验的比值：

$$w(Z) = \frac{P_\theta(X, Z)}{q_\phi(Z|X)}$$

对数重要性权重即为：

$$\log w(Z) = \log P_\theta(X, Z) - \log q_\phi(Z|X)$$

对数重要性权重方差

$$\mathrm{Var}_{Z\sim q_\phi}[\log w(Z)] \tag{10-7}$$

用以评估推断网络的质量：

- **方差越小**：说明 q_ϕ 作为重要性采样分布的质量越高，它越贴近真实的后验分布。
- **方差越大**：说明 q_ϕ 的拟合出现了严重偏差（例如 PoE 导致的方差塌陷），网络在隐空间中进行了低效甚至错误的采样。

分析：根据贝叶斯公式，联合分布可以写为 $P_\theta(X,Z)=P_\theta(Z|X)P(X)$。我们将其代入对数权重公式中：

$$\begin{aligned}\log w(Z) &= \log\left(P_\theta(Z|X)P(X)\right) - \log q_\phi(Z|X)\\ &= \log P(X) + \log\frac{P_\theta(Z|X)}{q_\phi(Z|X)}\end{aligned} \tag{10-8}$$

因此，若推断网络 $q_\phi(Z|X)$ 训练到至善，则其完全拟合真实的后验分布 $P_\theta(Z|X)$，那么 $\frac{P_\theta(Z|X)}{q_\phi(Z|X)}\to 1$，其对数值 $\to 0$。此时，对数重要性权重将完全退化为一个常数：

$$\log w(Z)\to \log P(X) \tag{10-9}$$

既然 $\log P(X)$ 对于给定的数据 X 是一个与 Z 毫无关系的常数，那么当我们从 q_ϕ 中多次采样不同的 Z 时，计算出的 $\log w(Z)$ 应该是一成不变的。换言之，如果近似后验是完美的，重要性权重的方差必须严格等于 0（$\mathrm{Var}[\log w]=0$）。

表 1: Average variance of log importance weights for three marginal probabilities, estimated by importance sampling from $q(z|x_1)$. 1000 importance samples were used.

——*MVAE, Wu & Goodman, 2018*

Model	BinaryMNIST	MNIST	FashionMNIST	MultiMNIST	CelebA
Variance of Marginal Log Importance Weights: var $\left(\log\frac{p(x_1,z)}{q(z\|x_1)}\right)$					
VAE	22.264	26.904	25.795	54.554	56.291
BiVCCA	55.846	93.885	33.930	185.709	429.045
JMVAE	39.427	37.479	53.697	84.186	331.865
MVAE-Q	34.300	37.463	34.285	69.099	100.072
MVAE	**22.181**	**25.640**	**20.309**	**26.917**	73.923
MVAE19	–	–	–	–	71.640
Variance of Joint Log Importance Weights: var $\left(\log\frac{p(x_1,x_2,z)}{q(z\|x_1)}\right)$					
JMVAE	41.003	40.126	56.640	91.850	334.887
MVAE-Q	34.615	38.190	34.908	64.556	101.238
MVAE	**23.343**	**27.570**	**20.587**	**27.989**	76.938
MVAE19	–	–	–	–	72.030
Variance of Conditional Log Importance Weights: var $\left(\log\frac{p(x_1,z\|x_2)}{q(z\|x_1)}\right)$					
CVAE	21.203	22.486	12.748	–	56.852
JMVAE	23.877	26.695	26.658	37.726	81.190
MVAE-Q	34.719	38.090	34.978	44.269	101.223
MVAE	**19.478**	25.899	18.443	**16.822**	73.885
MVAE19	–	–	–	–	71.824

10.4.2 定性评价：条件生成的视觉质量与语义对齐

"Fig. 2 shows image samples and conditional image samples for each dataset using the image generative model. We find the samples to be good quality, and find conditional samples to be largely correctly matched to the target label." ——*Wu & Goodman, MVAE (2018)*

对联合推断网络 q_ϕ 和双路解码器 P_θ 的双重验收：

- **样本质量**：检验模型是否发生了“模式崩溃 (Mode Collapse)”或生成了模糊的均值图像。这证明了即使在多模态妥协的压力下，图像解码器 $P_{\theta_{img}}(X|Z)$ 依然保持了极高的生成保真度。
- **条件匹配**：这是对**跨模态推理能力**最核心的定性检验。例如，给定一个离散的文本标签 X_2（如“短靴 (Ankle boot)”），模型必须先通过文本专家网络 $q_{\phi_2}(Z|X_2)$ 将该离散语义精准地映射到连续的隐空间 Z 中，随后只凭这个 Z，通过图像解码器 $P_{\theta_1}(\hat{X}_1|Z)$ 渲染出高维像素。如果生成的图像确实精准呈现了目标属性，这就从直观上证明了：模型成功在 Z 空间打通了异构数据（如文本与图像、基因与表型）之间的语义壁垒。

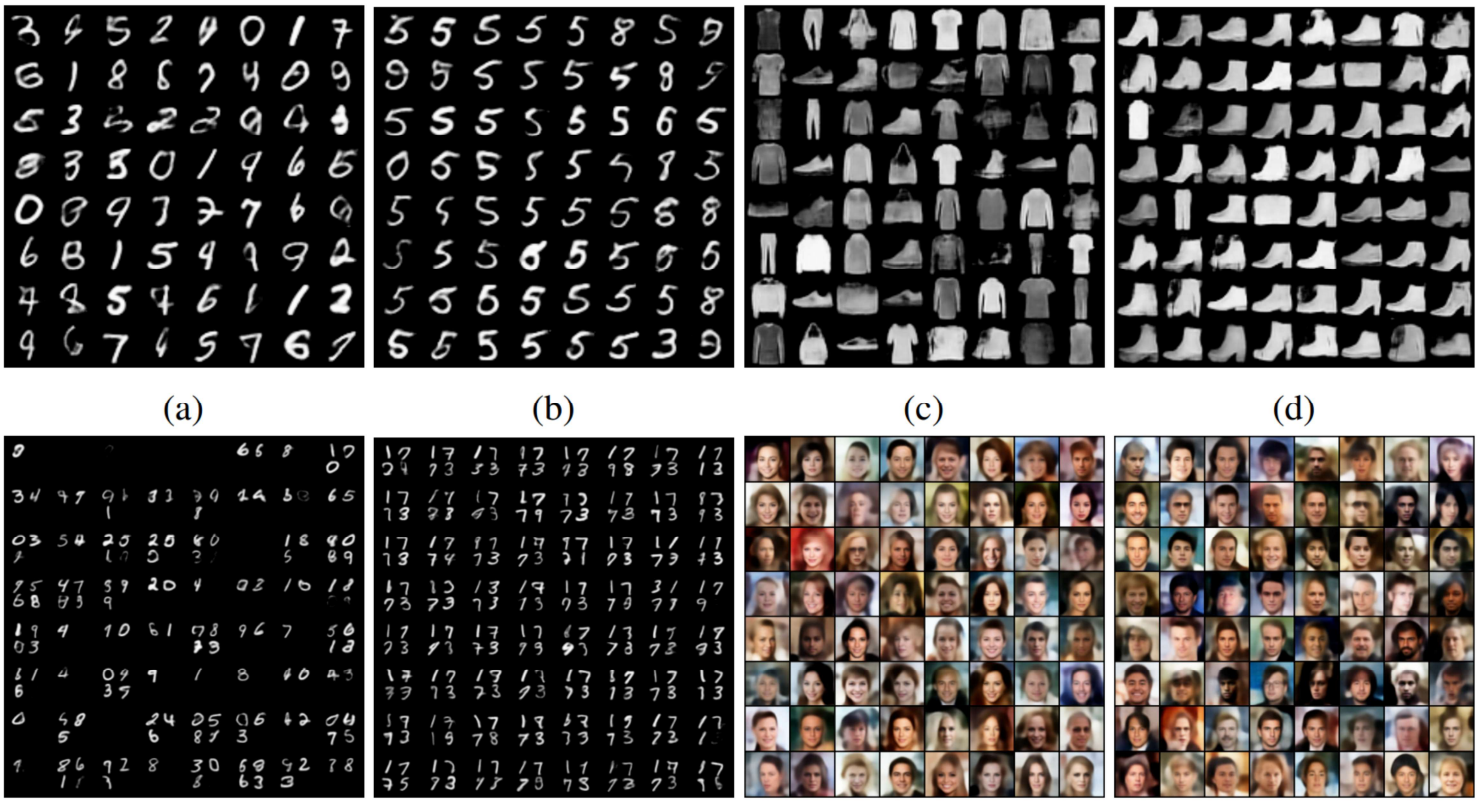

图 14: 使用 MVAE 的图像采样结果。(a, c, e, g) 展示了每个数据集的 64 张图像，其生成过程为：从先验分布 $z \sim p(z)$ 中采样，随后通过 $p(x_1|z)$ 进行解码生成。类似地，(b, d, f, h) 展示了条件图像的重建结果，其生成过程为从单模态推断网络 $z \sim q(z|x_2)$ 中采样，其中给定的目标条件标签分别为：(b) $x_2 = 5$，(d) $x_2 =$ 短靴 (Ankle boot)，(f) $x_2 = 1773$，(h) $x_2 =$ 男性 (Male)。

—— *Wu & Goodman, MVAE (2018)*

10.5 MVAE 过度乐观的“独立性假设”、“方差塌陷”，以及 HZAU-VAE 动态门控网络的新设计

10.5.1 今天的动机

MVAE 中的专家乘积 (PoE) 能够成立，建立在一个非常强硬的数学假设上：在给定隐变量 Z 的情况下，各个单模态提供的“证据”是完全独立的。但在真实的复杂物理与生物系统中，这种假设是“过度乐观 (Overoptimistic)”的。例如，当我们将高维的单细胞表达谱 X_1 与复杂的致癌文本语义 X_2 对齐时，这两个模态之间必然存在极强的底层非线性耦合。如果强行假设它们独立并进行简单的概率相加（即 PoE 的精度直接相加），会导致联合推断的方差被**严重低估（过度自信）**，进而让 ELBO 的计算偏离真实的下界。

此时，传统 PoE（无门控）的逻辑类似于“谁嗓门大听谁的”。PoE 纯粹依赖专家自己计算出的方差（内部自信度）。如果一个专家过度自信（方差算得很小，但其实预测是错的），整个模型就会被它带偏，这就是“方差塌陷”。

为了解决这个问题，我们（课堂同学们）尝试抛弃单模态独立假设，引入了显式建模模态依赖关系的**动态门控共识聚合机制**，转而显式地学习和利用各个模态子集之间的“依赖关系”。

HZAU-VAE（引入门控）的逻辑：引入了“外部的客观监督”。门控网络 $\mathbf{W}_\psi(X)$ 跳出了单模态的局限，它站在全局视角（看到所有的 X），客观地评判每个专家在当前样本上的真实能力。专家即使内部很自信，但如果门控网络认为当前场景下该模态不可靠，依然会强行关小它的门（降低 w_S）。

10.5.2 生成阶段的 PGM：与 VAE 同

在生成阶段，HZAU-VAE 依然保留了条件独立性假设 $P_\theta(X_1, X_2|Z) = P_{\theta_1}(X_1|Z)P_{\theta_2}(X_2|Z)$，以确保隐空间 Z 能够完全解耦和吸收多模态信息。

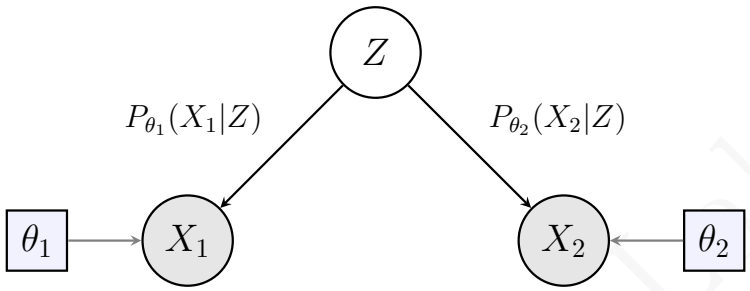

图 15: MVAE 与 HZAU-VAE 共用的生成概率图模型 (PGM)。在生成阶段，两者均保留了条件独立性假设：隐变量 Z 充当“公共原因”(Common Cause)，一旦 Z 给定，多模态特征的解码路径将被完全解耦，由各自独立的网络 θ_1 和 θ_2 负责。

10.5.3 推断阶段的 PGM：从“直接对撞”到“共识博弈”

模型的核心变化发生在**推断阶段**。与 MVAE 简单的“对撞节点”不同，HZAU-VAE 在推断 PGM 中引入了一个全新的“共识算子”网络 W_ψ。

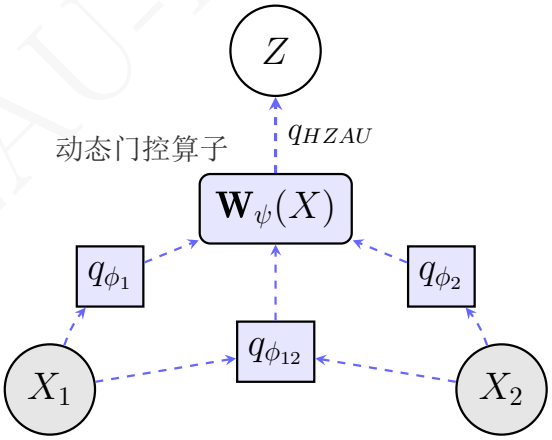

图 16: HZAU-VAE 的推断概率图模型 (PGM)。与 PoE 中专家之间互不干涉不同，HZAU-VAE 显式引入了联合子集专家 ($q_{\phi_{12}}$) 来捕捉模态间的底层依赖关系。最终，包含独立专家 (q_{ϕ_1}, q_{ϕ_2}) 在内的所有推断信息，交由参数化的共识算子 $\mathbf{W}_\psi$ 动态分配权重，共同输出联合后验 q_{HZAU}。

10.5.4 动态门控共识推断的解析公式

在数学层面，HZAU-VAE 机制放弃了简单的精度连加，转而采用一种更具泛化性的**动态加权对数池化 (Dynamically Weighted Log-Linear Pooling)**。

假设我们有模态子集集合 $\mathcal{S} = \{\{1\},\{2\},\{1,2\}\}$。模型会为每一个子集 $S \in \mathcal{S}$ 训练一个专家网络 $q_{\phi_S}(\mathbf{Z}|\mathbf{X}_S)$。同时，引入一个权重网络 $\mathbf{W}_\psi(\mathbf{X}_{1:M})$，用于输出各个专家在当前数据样本下的依赖权重 w_S（满足 $\sum w_S = 1$）。联合推断公式被重新定义为：[18]

$$q_{HZAU}(\mathbf{Z}|\mathbf{X}_1,\mathbf{X}_2) \propto p(\mathbf{Z}) \prod_{S\in\mathcal{S}} \left(\frac{q_{\phi_S}(\mathbf{Z}|\mathbf{X}_S)}{p(\mathbf{Z})}\right)^{w_S(\mathbf{X}_1,\mathbf{X}_2;\psi)} \tag{10-10}$$

定理 10-2. *(HZAU-VAE 的闭式解析解)* 若所有专家分布均服从对角高斯分布 $\mathcal{N}(\boldsymbol{\mu}, diag(\boldsymbol{\sigma}^2))$，且门控权重满足归一化约束 $\sum_{S\in\mathcal{S}} w_S = 1$。通过展开对数池化公式并消去先验项，*HZAU* 的联合后验分布依然为对角高斯分布，其参数具有如下极简闭式解：

$$\begin{cases} \dfrac{1}{\boldsymbol{\sigma}^2_{HZAU}} &= \displaystyle\sum_{S\in\mathcal{S}} \frac{w_S}{\boldsymbol{\sigma}^2_S} \\ \boldsymbol{\mu}_{HZAU} &= \boldsymbol{\sigma}^2_{HZAU} \odot \left(\displaystyle\sum_{S\in\mathcal{S}} w_S \frac{\boldsymbol{\mu}_S}{\boldsymbol{\sigma}^2_S}\right) \end{cases} \tag{10-11}$$

物理意义：在这个化简后的形态中，先验分布不再直接参与解析计算，而是通过影响单个专家的推断结果间接发挥作用。联合精度（方差的倒数）被定义为各专家精度的加权和；联合均值则是各专家均值的加权平均（以各自精度为权）。这种形式在 *PyTorch* 或 *TensorFlow* 实现中极具优势，仅需几行矩阵乘法即可完成高维特征的动态融合，完美规避了 *PoE* 的方差塌陷风险。

【定理证明】

如同标准 PoE 的推导，由于对角高斯分布的各维度相互独立，我们只需在单一维度的标量空间 z 下进行推导，最终结果可自然推广至多维空间。

根据 HZAU 的动态加权对数池化公式，我们将联合推断分布 $q_{HZAU}(z)$ 的各项展开：

$$\begin{aligned} q_{HZAU}(z) &\propto p(z) \prod_{S\in\mathcal{S}} \left(\frac{q_{\phi_S}(z)}{p(z)}\right)^{w_S} \\ &= p(z) \cdot \frac{\prod_{S\in\mathcal{S}} \left(q_{\phi_S}(z)\right)^{w_S}}{\prod_{S\in\mathcal{S}} \left(p(z)\right)^{w_S}} \\ &= p(z) \cdot \frac{\prod_{S\in\mathcal{S}} \left(q_{\phi_S}(z)\right)^{w_S}}{p(z)^{\sum_{S\in\mathcal{S}} w_S}} \end{aligned}$$

此时，引入门控网络的核心约束条件：所有子集的权重之和为 1，即 $\sum_{S\in\mathcal{S}} w_S = 1$。代入上式后，分母变为 $p(z)^1 = p(z)$，它与最外面的先验 $p(z)$ 完美抵消！因此，联合分布极其优雅地化简为所有专家分布的加权几何平均：

$$q_{HZAU}(z) \propto \prod_{S\in\mathcal{S}} \left(q_{\phi_S}(z)\right)^{w_S}$$

[18] 与 PoE 公式相比，HZAU-VAE 为每个专家项引入了由神经网络 ψ 动态计算的非线性指数权重 w_S。当专家之间存在高度重叠（强依赖）时，门控网络会自动降低它们的权重，防止“无效证据”被重复计算而导致方差过小。同时，如果我们比较 (eq. 10-3)，

$$P(Z|X_1,X_2) \propto \frac{P(Z|X_1)}{P(Z)} \frac{P(Z|X_2)}{P(Z)} P(Z),$$

HZAU-VAE 的推断拆分看上去更贴近 Z 的理论后验的分解。

接下来，我们将单维高斯分布的概率密度函数 $q_{\phi_S}(z) \propto \exp\left(-\frac{(z-\mu_S)^2}{2\sigma_S^2}\right)$ 代入：

$$\begin{aligned} q_{HZAU}(z) &\propto \prod_{S\in\mathcal{S}} \exp\left(-w_S \frac{(z-\mu_S)^2}{2\sigma_S^2}\right) \\ &= \exp\left(-\frac{1}{2}\sum_{S\in\mathcal{S}} w_S \frac{(z-\mu_S)^2}{\sigma_S^2}\right) \end{aligned}$$

将指数项内部的二次型完全展开，并按 z 的幂次合并同类项：

$$\begin{aligned} \sum_{S\in\mathcal{S}} w_S \frac{(z-\mu_S)^2}{\sigma_S^2} &= \sum_{S\in\mathcal{S}} \left(\frac{w_S}{\sigma_S^2} z^2 - \frac{2w_S\mu_S}{\sigma_S^2} z + \frac{w_S\mu_S^2}{\sigma_S^2}\right) \\ &= \left(\sum_{S\in\mathcal{S}} \frac{w_S}{\sigma_S^2}\right) z^2 - 2\left(\sum_{S\in\mathcal{S}} \frac{w_S\mu_S}{\sigma_S^2}\right) z + \sum_{S\in\mathcal{S}} \frac{w_S\mu_S^2}{\sigma_S^2} \end{aligned}$$

我们期望最终的 $q_{HZAU}(z)$ 是一个高斯分布 $\mathcal{N}(\mu_{HZAU}, \sigma_{HZAU}^2)$，其标准指数展开式为：

$$\frac{(z-\mu_{HZAU})^2}{\sigma_{HZAU}^2} = \frac{1}{\sigma_{HZAU}^2} z^2 - \frac{2\mu_{HZAU}}{\sigma_{HZAU}^2} z + \frac{\mu_{HZAU}^2}{\sigma_{HZAU}^2}$$

通过对比系数（系数匹配法），我们立刻得到：

$$\begin{cases} \dfrac{1}{\sigma_{HZAU}^2} = \displaystyle\sum_{S\in\mathcal{S}} \frac{w_S}{\sigma_S^2} & (\text{对应} z^2 \text{ 的系数}) \\ \dfrac{\mu_{HZAU}}{\sigma_{HZAU}^2} = \displaystyle\sum_{S\in\mathcal{S}} \frac{w_S\mu_S}{\sigma_S^2} & (\text{对应} z \text{ 的系数}) \end{cases}$$

将第一式的结论代入第二式，移项即可得到联合均值 μ_{HZAU}。由于多维独立，我们将上述一维标量的结论推广至每个维度 $d \in \{1, \dots, D\}$，写成向量的逐元素运算形式 $\odot$，即得证原定理。

10.6 HZAU-VAE 的变分下界 (ELBO)

通过将全新的联合后验 $q_{HZAU}(\mathbf{Z}|\mathbf{X}_1, \mathbf{X}_2)$ 注入标准变分推导，我们得到了 HZAU-VAE 的优化目标函数。其展开形式在宏观上与 MVAE 保持一致，但在推断项 q 上发生了质的替换。

我们用 HZAU-VAE 门控网络替换掉原有的 PoE 公式，得到优化联合对数似然的 **HZAU-ELBO**：

$$\begin{aligned} \mathcal{L}_{HZAU}(\mathbf{X}_1, \mathbf{X}_2) &= \mathbb{E}_{q_{HZAU}(\mathbf{Z}|\mathbf{X}_1,\mathbf{X}_2)}\left[\log \frac{P_\theta(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Z})}{q_{HZAU}(\mathbf{Z}|\mathbf{X}_1, \mathbf{X}_2)}\right] \\ &= \underbrace{\mathbb{E}_{q_{HZAU}}\left[\log P_{\theta_1}(\mathbf{X}_1|\mathbf{Z}) + \log P_{\theta_2}(\mathbf{X}_2|\mathbf{Z})\right]}_{\text{共识驱动的双路重构似然 (MSE/BCE)}} - \underbrace{\mathcal{D}_{KL}(q_{HZAU}(\mathbf{Z}|\mathbf{X}_1, \mathbf{X}_2) \,\|\, p(\mathbf{Z}))}_{\text{共识后验与先验的 KL 惩罚}} \end{aligned} \tag{10-12}$$

在实际的深度学习框架（如 PyTorch）中，我们需要最大化 ELBO，这等价于最小化负的 ELBO（即最终的 Loss 函数）。我们将 HZAU-ELBO 的两项拆解来看：

第一项：重构似然的计算 (基于重参数化技巧的蒙特卡洛采样)

公式中的 $\mathbb{E}_{q_{HZAU}}[\dots]$ 是一个期望操作。由于解码器网络 P_θ 是高度非线性的，我们无法求出这个期望的解析解。工程上的标准做法是使用**单样本蒙特卡洛估计 (Single-sample Monte Carlo Estimation)** 结合**重参数化技巧 (Reparameterization Trick)**。

计算步骤如下：

1. **融合参数**：首先通过前向传播，利用我们在定理中推导的闭式解，计算出联合分布的参数 $\boldsymbol{\mu}_{HZAU}$ 和 $\boldsymbol{\sigma}_{HZAU}^2$。

2. **重参数化采样：**为了让梯度能够穿透随机采样过程流回编码器和门控网络，我们从标准正态分布中采样噪声 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，然后通过仿射变换得到隐变量 $\mathbf{z}$：

$$\mathbf{z} = \boldsymbol{\mu}_{HZAU} + \boldsymbol{\sigma}_{HZAU} \odot \boldsymbol{\epsilon} \tag{10-13}$$

3. **计算损失：**将采样得到的这一组 $\mathbf{z}$ 送入两个解码器，得到重构结果 $\hat{\mathbf{X}}_1$ 和 $\hat{\mathbf{X}}_2$。此时，最大化对数似然在代码中等价于计算重构误差（MSE 损失）：

$$\mathcal{L}_{Recon} = \text{Loss}_{func}(\mathbf{X}_1, \hat{\mathbf{X}}_1) + \text{Loss}_{func}(\mathbf{X}_2, \hat{\mathbf{X}}_2) \tag{10-14}$$

第二项：KL 散度的计算 (拥有完美闭式解)

与重构项不同，由于我们在 HZAU-VAE 中刻意保持了所有的分布都是对角高斯分布，并且先验 $p(\mathbf{Z})$ 被设定为标准正态分布 $\mathcal{N}(\mathbf{0}, \mathbf{I})$。这两个对角高斯分布之间的 KL 散度**不需要任何采样，它拥有绝对精确的解析解** (eq. 6-6)。[19]

假设隐空间 $\mathbf{Z}$ 共有 D 维，对于第 d 个维度，其 KL 散度的标量计算公式为：

$$\mathcal{D}_{KL}(q^{(d)}_{HZAU} \parallel p^{(d)}) = -\frac{1}{2}\left(1 + \log(\sigma^2_{HZAU,d}) - \mu^2_{HZAU,d} - \sigma^2_{HZAU,d}\right) \tag{10-15}$$

在向量化编程中，我们可以直接对整个 D 维向量进行操作，并对其求和：

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum_{d=1}^{D}\left(1 + \log(\boldsymbol{\sigma}^2_{HZAU}) - \boldsymbol{\mu}^2_{HZAU} - \boldsymbol{\sigma}^2_{HZAU}\right)_d \tag{10-16}$$

---

[19]根据信息论中两个 d 维多元高斯分布 $p = \mathcal{N}(\mu_1, \Sigma_1)$ 与 $q = \mathcal{N}(\mu_2, \Sigma_2)$ 之间 KL 散度的通用计算公式：

$$D_{KL}(p \parallel q) = \frac{1}{2}\left[\text{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1) - d + \log\frac{|\Sigma_2|}{|\Sigma_1|}\right]$$

在 HZAU-VAE 中，我们要计算的是近似后验 $q_{HZAU}(\mathbf{Z}|\mathbf{X})$ 与先验 $p(\mathbf{Z})$ 之间的 KL 散度。为了推导方便，我们省略下标，设后验为 q，先验为 p。

根据前文设定，近似后验 q 为对角高斯分布：$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。其中均值 $\mu_1 = \boldsymbol{\mu}_{HZAU}$，协方差矩阵 $\Sigma_1 = \text{diag}(\boldsymbol{\sigma}^2_{HZAU})$。先验 p 为标准多变量正态分布：$\mathcal{N}(\mathbf{0}, \mathbf{I})$。其中均值 $\mu_2 = \mathbf{0}$，协方差矩阵 $\Sigma_2 = \mathbf{I}$（单位矩阵）。

代入通用 KL 散度公式（设隐空间总维度为 D）：

1. 迹项 (Trace Term)：由于 $\Sigma_2 = \mathbf{I}$，其逆矩阵同样为单位矩阵 $\Sigma_2^{-1} = \mathbf{I}$。任何矩阵与单位矩阵相乘保持不变，且对角矩阵的迹（Trace）等于其主对角线元素之和。

$$\text{tr}(\Sigma_2^{-1}\Sigma_1) = \text{tr}(\mathbf{I} \cdot \text{diag}(\boldsymbol{\sigma}^2)) = \text{tr}(\text{diag}(\boldsymbol{\sigma}^2)) = \sum_{d=1}^{D}\sigma_d^2$$

2. 二次型项 (Quadratic Term)：将 $\mu_2 = \mathbf{0}$ 代入：

$$(\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1) = (\mathbf{0} - \boldsymbol{\mu})^T\mathbf{I}(\mathbf{0} - \boldsymbol{\mu}) = (-\boldsymbol{\mu})^T(-\boldsymbol{\mu}) = \boldsymbol{\mu}^T\boldsymbol{\mu} = \sum_{d=1}^{D}\mu_d^2$$

3. 常数项：由推导前文可知，该项即为总维度大小 D。

4. 行列式对数项 (Log-Determinant Term)：单位矩阵的行列式 $|\Sigma_2| = |\mathbf{I}| = 1$。而对角矩阵的行列式为其对角线元素的连乘积，即 $|\Sigma_1| = \prod_{d=1}^{D}\sigma_d^2$。利用对数的除法与乘法性质：

$$\log\frac{|\Sigma_2|}{|\Sigma_1|} = \log\frac{1}{\prod_{d=1}^{D}\sigma_d^2} = \log(1) - \log\left(\prod_{d=1}^{D}\sigma_d^2\right) = 0 - \sum_{d=1}^{D}\log(\sigma_d^2) = -\sum_{d=1}^{D}\log(\sigma_d^2)$$

汇总合并：将上述四项的化简结果重新代回原公式，并在最外部提取出一个负号，即可得到最终的解析解：

$$\mathcal{D}_{KL}(q \parallel p) = \frac{1}{2}\left[\sum_{d=1}^{D}\sigma_d^2 + \sum_{d=1}^{D}\mu_d^2 - D - \sum_{d=1}^{D}\log(\sigma_d^2)\right] = \frac{1}{2}\sum_{d=1}^{D}\left[\sigma_d^2 + \mu_d^2 - 1 - \log(\sigma_d^2)\right] = -\frac{1}{2}\sum_{d=1}^{D}\left[1 + \log(\sigma_d^2) - \mu_d^2 - \sigma_d^2\right]$$

总结：HZAU-VAE 最终优化的工程损失函数

将上述两项结合，在实际的优化器（如 Adam）中，我们要最小化的总体目标损失为：

$$\text{Loss}_{total} = \underbrace{\lambda_1 \text{MSE}(\mathbf{X}_1, \hat{\mathbf{X}}_1) + \lambda_2 \text{BCE}(\mathbf{X}_2, \hat{\mathbf{X}}_2)}_{\text{最大化期望对数似然 (带权重平衡)}} + \beta \underbrace{\mathcal{L}_{KL}(\boldsymbol{\mu}_{HZAU}, \boldsymbol{\sigma}^2_{HZAU})}_{\text{最小化 KL 散度}}$$

注：由于门控权重 w_S 参与了 $\boldsymbol{\mu}_{HZAU}$ 和 $\boldsymbol{\sigma}^2_{HZAU}$ 的计算，因此在反向传播时，$Loss_{total}$ 的梯度会自然地流向门控网络 $\mathbf{W}_\psi$，引导它学会“哪个专家的重构能力更强，就给谁更高的权重”。

10.7 HZAU-VAE 的代码实施

10.7.1 门控网络 $\mathbf{W}_\psi$ 的结构设计

通常，门控算子 $\mathbf{W}_\psi$ 会被实现为一个轻量级的多层感知机（MLP）。为了保证前文推导中 $\sum w_S = 1$ 的权重归一化约束，其输出层必须应用 `Softmax` 激活函数。

```
i
```

```
mport torch
import torch.nn as nn
import torch.nn.functional as F

class GatingNetwork(nn.Module):
    def __init__(self, input_dim, num_experts):
        super(GatingNetwork, self).__init__()
        self.net = nn.Sequential(
            nn.Linear(input_dim, 128),
            nn.ReLU(),
            nn.Linear(128, num_experts),
            nn.Softmax(dim=-1) # 核心：确保输出的依赖权重之和为 1
        )

    def forward(self, x):
        # x 通常为各模态观测数据或其底层特征(Embedding)的拼接
        return self.net(x)
```

10.7.2 HZAU-VAE 的核心前向传播 (Forward) 流程

在实现网络的前向传播（`forward`）函数时，整个推断与生成逻辑可以清晰地划分为以下核心阶段：

1. **并行推断 (Parallel Inference)**：每个专家网络（包括各单模态编码器与联合编码器）独立前向计算，输出各自后验分布的参数 $\boldsymbol{\mu}_S$ 和对数方差 $\log \boldsymbol{\sigma}^2_S$。
2. **动态权重计算 (Dynamic Weight Allocation)**：门控网络接收模态拼接特征，动态输出分配给各个专家的归一化权重 w_S。
3. **门控融合 (Closed-form Fusion)**：严格利用我们定理中推导出的极简闭式解，执行加权精度与加权均值的张量运算，得出联合后验的参数 $\boldsymbol{\mu}_{HZAU}$ 和 $\boldsymbol{\sigma}^2_{HZAU}$。

注：融合完成后，即可通过标准的重参数化技巧进行采样，并送入解码器完成双路重构。

```c
lass HZAU_VAE(nn.Module):
    def __init__(self):
        super().__init__()
        # 定义专家编码器 (例如：图像专家、文本专家、联合专家)
        self.experts = nn.ModuleList([Encoder1(), Encoder2(), Encoder12()])
        # 定义门控网络 (输入为两模态拼接，输出3个权重)
        self.gating_net = GatingNetwork(input_dim_x1 + input_dim_x2, num_experts=3)

    def forward(self, x1, x2):
        # 1. 并行推断：获取各专家预测的分布参数 (D维向量)
        # mu_list 形状: [num_experts, batch_size, latent_dim]
        mu_list, logvar_list = [], []
        inputs = [x1, x2, torch.cat([x1, x2], dim=-1)] # 对应三个专家的输入
        for i, encoder in enumerate(self.experts):
            mu, logvar = encoder(inputs[i])
            mu_list.append(mu)
            logvar_list.append(logvar)

        # 2. 计算动态权重 w_S
        # weights 形状: [batch_size, num_experts]
        combined_x = torch.cat([x1, x2], dim=-1)
        weights = self.gating_net(combined_x)

        # 3. 门控融合 (计算加权精度和加权均值)
        # 将权重扩展维度以便广播运算: [batch_size, num_experts, 1]
        w = weights.unsqueeze(-1)

        # 计算精度 1/sigma^2 (利用 logvar 防止数值溢出)
        precisions = torch.stack([torch.exp(-lv) for lv in logvar_list], dim=1)
        var_hzau = 1.0 / torch.sum(w * precisions, dim=1)

        # 计算加权均值 mu
        mus = torch.stack(mu_list, dim=1)
        mu_hzau = var_hzau * torch.sum(w * precisions * mus, dim=1)

        # 4. 重参数化采样 (z = mu + sigma * epsilon)
        z = self.reparameterize(mu_hzau, var_hzau)

        # 5. 解码重建
        recon_x1 = self.decoder1(z)
        recon_x2 = self.decoder2(z)

        return recon_x1, recon_x2, mu_hzau, var_hzau
```

10.8 总结

这个算法的设计，标志着 HZAU《数据挖掘》课堂的同学能掌握隐变量模型的设计和求解。在这个场景中，对于涉及多源异构数据联合表征的架构而言，直接套用传统的 MVAE 可能会因为模态间的强依赖关系而陷入潜在的推断困境。HZAU-VAE 的设计尝试证明：只要在 ELBO 的计算源头上重新设计变分后验的分解公式，赋予模型理解“模态间依赖贡献度”的能力，就有望在适量增加模型体积的前提下，提升跨模态映射的逼真度与潜在特征的纯度。