Méthodes de simulation : Implémentation du VAEAC et analyse de l'impact du prior gaussien sur l'explicabilité

OULAD ALI Ayoub

Sorbonne Université – MS2A
ayoubouladali30@gmail.com

3 février 2026

Résumé

L'essor des modèles prédictifs complexes, souvent qualifiés de « boîtes noires », a rendu l'explicabilité (XAI) indispensable. Au-delà du simple besoin de transparence, l'enjeu est désormais critique en production : il s'agit de détecter les biais et de comprendre les décisions même lorsque la distribution des données évolue. Ce projet explore une réponse statistique à ce défi en s'appuyant sur les travaux d'Olsen et al. (2022) : comment calculer des valeurs de Shapley justes lorsque les variables sont interdépendantes ? Ignorer ces corrélations revient en effet à baser les explications sur des combinaisons de données aberrantes, faussant ainsi l'interprétation.Nous analysons ici la solution du VAEAC (Variational Autoencoder with Arbitrary Conditioning), un modèle génératif conçu pour reconstruire n'importe quel sous-ensemble de variables manquantes en respectant la structure des données. Après avoir détaillé son architecture et ses mécanismes clés (rôle des masques, objectif ELBO, skip-connections), nous évaluons sa capacité à modéliser les distributions conditionnelles $p(x_{\bar{S}}|x_S)$ sur des données réelles et synthétiques.Notre analyse met en lumière deux limitations structurelles majeures : un coût de calcul élevé dû à l'échantillonnage Monte Carlo, et une tendance au « lissage » face à des distributions multimodales. Nous démontrons que ces défauts ne sont pas anodins : une mauvaise imputation biaise l'estimation de l'espérance conditionnelle, conduisant mécaniquement à des valeurs de Shapley erronées et, in fine, à des explications potentiellement trompeuses

Contributions analyse.

- Analyse unifiée du VAEAC pour l'imputation arbitraire et l'estimation de $\mathbb{E}[f(x) \mid x_S]$ (cadre SHAP conditionnel).
- Mise en évidence et explication théorique du *lissage* sur distributions multimodales.
- Passage du VAEAC au Diffusion with abstract conditionning
- Tests multiples (en général 3 à 5 reruns par test) sur différents datasets réels ou simulés, analyse des limites du modèle et des conséquences.

Table des matières

Table des figures **3**

1 Cadre Théorique : De la Théorie des Jeux à l'Explicabilité **5**
1.1 Théorie des jeux coopératifs . . . 5
1.2 Valeurs de Shapley . . . 5
1.2.1 Propriétés des Shapley values . . . 5
1.3 Application à l'Explicabilité (SHAP) . . . 6

2 Auto-Encodeurs Variationnels et Conditionnement **8**
2.1 VAE (Variational Autoencoder) . . . 8
2.2 VAEAC : VAE with Arbitrary Conditioning . . . 8
2.2.1 Architecture du VAEAC : Deux réseaux principaux . . . 9
2.2.2 Objectif d'optimisation et ELBO conditionnelle . . . 10
2.3 Rôle Structurel du Masque d'Observation . . . 11
2.4 Positionnement par rapport aux architectures standards . . . 11
2.5 Traitement des Données Manquantes : Deux Protocoles Distincts . . . 12
2.5.1 Protocole 1 : VAEAC Standard (Marginalisation) . . . 12
2.6 Métriques d'Évaluation . . . 12
2.6.1 Métriques pour Valeurs de Shapley . . . 13
2.6.2 Métriques de Reconstruction Auxiliaires . . . 13
2.7 Fonction de Coût et Régularisation . . . 13

3 Analyse de la Qualité de Reconstruction **14**
3.1 Résultats Expérimentaux . . . 14
3.2 Analyse de la Fidélité Explicative . . . 14
3.3 Impact de la Dimension du Masque . . . 15

4 Analyse de Sensibilité : Impact de ϵ **16**
4.1 Visualisation des Performances . . . 16

5 Stress-Test de Robustesse : Censure MNAR **17**
5.1 Protocole de Censure et Biais . . . 17
5.2 Résultats Quantitatifs . . . 17
5.3 Analyse : L'Échec de l'Extrapolation face à la Stabilité Explicative . . . 17

6 Analyse Théorique : Le problème du Lissage **18**
6.1 Cadre simplifié pour l'analyse . . . 18
6.1.1 Préservation de la symétrie et de la multimodalité par conditionnement . . . 19
6.1.2 Vecteurs gaussiens quelconques . . . 19
6.2 Preuve : La Minimisation L^2 induit le Lissage Conditionnel . . . 20
6.3 Peut-on éviter le flou ? . . . 21
6.3.1 Succès de MissForest : Partitionnement vs Continuité . . . 21
6.4 Proposition de Solution : Régularisation par Ancrage Modal . . . 22
6.4.1 Formulation Mathématique . . . 22
6.4.2 Algorithme Généralisé et Sélection de Modèle . . . 23
6.4.3 Limitations et considérations théoriques . . . 24
6.5 Dérivation de l'ELBO du VAEAC . . . 26
6.5.1 Lien entre Log-Vraisemblance et MSE . . . 27

7 Analyses Complémentaires et Détails Techniques **27**
7.1 Architecture des Skip-Connections . . . 27
7.2 Analyse Mathématique des Skip-Connections : Résolution du Vanishing Gradient . . . 28
7.3 Analyse de la sensibilité aux choix des hyperparamètres . . . 30

A Cas général : Décodeur avec variance apprise **33**
A.1 Formulation du problème . . . 33
A.2 Solution optimale théorique . . . 33
A.3 Application au cas bimodal symétrique . . . 33
A.4 Interprétation . . . 33

B Du VAEAC au DAC : du VAE à la diffusion conditionnelle **34**
B.1 A.1 Point de départ : VAE et log-vraisemblance (ELBO) . . . 34
B.2 A.2 Évolution : modèles de diffusion comme "cascade de latents" . . . 34
B.3 A.3 Conditionnement arbitraire et imputation . . . 34
B.4 A.4 Diffusion avec conditionnement arbitraire (DAC) : Alignement Variationnel . . . 35
B.5 Analyse des Résultats . . . 35
B.6 Analyse Quantitative et Comparative . . . 35
B.6.1 La Suprématie de la Diffusion sur l'Information Dispersée (Masque Damier) . . . 35
B.6.2 Métrique de Qualité : Peak Signal-to-Noise Ratio (PSNR) . . . 35
B.6.3 L'Efficacité Inattendue de la Diffusion Rapide ($T = 5$ vs $T = 10$) . . . 36
B.6.4 Compromis Stabilité (VAEAC) vs Réalisme (DAC) . . . 36
B.6.5 Conclusion Partielle . . . 37

C Travaux connexes, limitations et perspectives **37**
C.0.1 Quelques test disponibles ici pour le reproductibilité des résultats et la vérification de l'architecture . . . 38

Table des figures

1 **Processus d'inférence du VAEAC.** Le modèle génère les échantillons manquants $x_{\bar{S}}$ conditionnellement aux observations x_S. Durant l'apprentissage, un "Proposal Network" (encodeur complet) guide l'apprentissage, tandis qu'au déploiement, seul le "Prior Network" (encodeur masqué) est utilisé pour générer les complétions via l'espace latent, tiré du papier original. . . . 9
2 **Schéma détaillé du VAEAC.** Comparaison des flux de données entre l'entraînement (où x complet est connu) et le test (où seul x_S est connu).image tiré du papier original. . . . 10
3 Évaluation de la fidélité des valeurs de Shapley (Moyenne ± Écart-type) selon la méthode d'imputation sur le dataset *Abalone*. Les résultats sont agrégés sur 3 itérations (*runs*) indépendantes. . . . 14
4 **Analyse des zones d'incertitude :** Évolution de l'erreur de reconstruction (RMSE) en fonction du nombre de caractéristiques masquées ($k \in \{1, 2, 4, 8\}$) pour les datasets IID, GMM et Copula. Le taux de données manquantes global est fixé à 30%. . . . 16
5 **Dynamique de l'Imputation :** Comparaison de l'évolution de la RMSE, du R^2 et de la Divergence KL pour MissForest et VAEAC. Les zones ombrées représentent l'intervalle de confiance (Std Dev). . . . 16
6 **Changement d'énergie du modèle.** Loss standard (L_2^2) et loss modifiée $L_2^2 + \lambda \min(\text{dist}^2)$. 23
7 **Validation de la Régularisation Modale.** La méthode parvient à ancrer les prédictions sur les modes réels plutôt que sur la moyenne (creux de densité), validant l'approche théorique. . . . 25
8 Évolution qualitative des reconstructions VAEAC en fonction de λ (Masque : Moitié Basse). 26

9 Architecture des Skip-Connections dans le VAEAC. Les activations de chaque couche de l'encodeur sont concaténées aux entrées symétriques du décodeur, préservant l'information locale tout en permettant une compression globale via l'espace latent z. 28
10 **Impact des Skip-Connections sur l'optimisation (échelle log).** À gauche : convergence de la loss ; au centre : flux de gradient moyen ; à droite : gradient minimal, comparant les modèles avec et sans skip-connections. 29
11 **Comparaison visuelle et métrique de l'imputation (VAEAC vs DAC).** Reconstruction du chiffre "8" sous trois scénarios de masquage (Vertical, Latéral, Damier). Les résultats comparent le VAEAC classique et le modèle DAC avec $T = 5$ et $T = 10$ pas de diffusion. *Détails expérimentaux :* Entraînement sur **10 époques** avec une taille de lot (*batch size*) de **64**. Ces résultats sont issus d'une **architecture générale légèrement modifiée** et utilisent une **fonction de perte adaptée**. Les scores RMSE (en rouge/vert) indiquent la qualité de reconstruction pixel à pixel. 37

1 Cadre Théorique : De la Théorie des Jeux à l'Explicabilité

1.1 Théorie des jeux coopératifs

Un jeu, au sens de la théorie des jeux, modélise l'interaction entre plusieurs agents. Soit $N = \{1, \dots, n\}$ un ensemble fini de joueurs. On appelle coalition tout sous-ensemble $C \subseteq N$.

Définition 1.1 (Jeu coopératif)**.** Un jeu est caractérisé par la paire (N, v) où $v : 2^N \to \mathbb{R}$ est la fonction caractéristique (ou fonction de paiement). Pour toute coalition C, $v(C)$ représente la valeur créée par la coopération des membres de C, avec par convention $v(\emptyset) = 0$.

La structure des coalitions est déterminante pour l'analyse. Une structure de coalition sur N est une partition $CS = \{C^1, \dots, C^k\}$ telle que $\bigcup_i C^i = N$ et $C^i \cap C^j = \emptyset$ pour $i \neq j$.

Deux propriétés fondamentales permettent de caractériser la dynamique de coopération :

1. **Monotonie :** Un jeu est monotone si l'ajout de joueurs ne diminue jamais la valeur de la coalition : $\forall C \subseteq D \subseteq N, v(C) \leq v(D)$.
2. **Super-additivité :** Si l'union de deux coalitions disjointes rapporte plus que la somme de leurs valeurs séparées ($v(C \cup D) \geq v(C) + v(D)$), le jeu incite à la "grande coalition" N.

Une condition plus forte est la *convexité*, définie par $v(C \cup D) + v(C \cap D) \geq v(C) + v(D)$, qui implique que la contribution marginale d'un joueur croît avec la taille de la coalition.

1.2 Valeurs de Shapley

La problématique centrale est l'allocation équitable de la valeur totale $v(N)$ entre les joueurs. La solution canonique, introduite par Lloyd Shapley (1953), repose sur l'idée que chaque joueur doit recevoir une rétribution proportionnelle à sa contribution marginale moyenne.

Définition 1.2 (Contribution marginale)**.** Soit Π_N l'ensemble des permutations de N. Pour une permutation $\pi \in \Pi_N$, on note $S_\pi(i)$ l'ensemble des prédécesseurs de i. La contribution marginale du joueur i est :

$$\Delta_\pi(i) = v(S_\pi(i) \cup \{i\}) - v(S_\pi(i)). \tag{1}$$

La valeur de Shapley $\phi_i(v)$ est alors définie comme l'espérance de la contribution marginale sur l'ensemble des permutations équiprobables :

$$\phi_i(v) = \frac{1}{n!} \sum_{\pi \in \Pi_N} \Delta_\pi(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left(v(S \cup \{i\}) - v(S)\right). \tag{2}$$

.

1.2.1 Propriétés des Shapley values

Les valeurs de Shapley sont l'unique solution satisfaisant un ensemble de quatre axiomes fondamentaux, garantissant une attribution équitable.

Propriété 1.1 (Axiomes de Shapley)**.** *Les valeurs de Shapley satisfont les propriétés suivantes :*

Efficacité (Efficiency) : *La somme des contributions est égale à la valeur de la grande coalition N moins la valeur de l'ensemble vide.*

$$\sum_{j \in N} \phi_j = v(N) - v(\emptyset)$$

,

Symétrie (Symmetry) : *Deux joueurs j et k qui contribuent de manière égale à toutes les coalitions reçoivent le même paiement. C'est-à-dire, si $v(S \cup \{j\}) = v(S \cup \{k\})$ pour tout $S \subseteq N \setminus \{j, k\}$, alors :*

$$\phi_j = \phi_k$$

,

Joueur Nul (Dummy) : *Un joueur j qui ne contribue pas, c'est-à-dire que sa présence ne change pas la valeur de la coalition ($v(S) = v(S \cup \{j\})$ pour tout S), reçoit un paiement nul :*

$$\phi_j = 0,$$

Linéarité (Linearity) : *Pour une combinaison linéaire de m jeux $\{v_1, ..., v_m\}$, définie par $v(S) = \sum_{k=1}^{m} c_k v_k(S)$, les valeurs de Shapley respectent cette linéarité :*

$$\phi_j(v) = \sum_{k=1}^{m} c_k \phi_j(v_k),$$

Définition 1.3 (Formule de Shapley)**.** La valeur de Shapley pour un joueur j, notée ϕ_j, est donnée par la moyenne pondérée des contributions marginales :

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(n - |S| - 1)!}{n!} \Big(v(S \cup \{j\}) - v(S) \Big),$$

où $n = |N|$ est le nombre total de joueurs, et la somme parcourt tous les sous-ensembles S de N qui ne contiennent pas j.

1.3 Application à l'Explicabilité (SHAP)

Dans le contexte de l'apprentissage automatique, nous transposons le formalisme de la théorie des jeux coopératifs au problème de l'interprétation des modèles prédictifs.

Contexte. Soit un jeu de données $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ où $x \in \mathbb{R}^d$. Pour expliquer une prédiction locale $\hat{y} = f(x^*)$ à l'aide des valeurs de Shapley, nous définissons :

- **Ensemble des joueurs** : $N = \{1, \ldots, d\}$ correspond aux d caractéristiques (features) de l'entrée x.
- **Fonction caractéristique** : Pour chaque coalition $S \subseteq N$, la valeur $v(S)$ représente la "performance attendue" du modèle lorsque seules les variables dans S sont observées.

Définition de la fonction de valeur. Lundberg et Lee (2017) proposent de définir la fonction caractéristique par l'espérance conditionnelle :

$$v(S) = \mathbb{E}[f(x) \mid x_S = x_S^*], \tag{3}$$

où :

- $x_S^* \in \mathbb{R}^{|S|}$ désigne les valeurs observées des features dans la coalition S pour l'instance x^* à expliquer
- $x_{\bar{S}}$ représente les variables non observées (complémentaires)
- L'espérance est prise sur la distribution marginale de $x_{\bar{S}}$ conditionnellement à $x_S = x_S^*$

On peut réécrire explicitement cette espérance comme :

$$v(S) = \mathbb{E}_{x_{\bar{S}} \sim p(\cdot \mid x_S^*)}[f(x_{\bar{S}}, x_S^*)]. \tag{4}$$

Défi computationnel : estimation de l'espérance conditionnelle. L'évaluation directe de $v(S)$ nécessite de calculer une intégrale sur toutes les valeurs possibles de $x_{\bar{S}}$, ce qui est généralement intraitable analytiquement. Une approximation de Monte Carlo classique s'écrit :

$$\hat{v}(S) \approx \frac{1}{K} \sum_{k=1}^{K} f(x_{\bar{S}}^{(k)}, x_S^*) \quad \text{avec} \quad x_{\bar{S}}^{(k)} \sim p(x_{\bar{S}} \mid x_S = x_S^*). \tag{5}$$

Le cœur du problème réside dans la modélisation fidèle de la distribution conditionnelle $p(x_{\bar{S}} \mid x_S)$:

— **Approches naïves** (ex : KernelSHAP) supposent l'indépendance des features : $p(x_{\bar{S}} \mid x_S) = p(x_{\bar{S}})$. Cette hypothèse est souvent irréaliste et conduit à échantillonner des combinaisons de features *hors distribution*, faussant ainsi l'estimation de $v(S)$.
— **Approches avancées** nécessitent un modèle génératif capable d'imputer $x_{\bar{S}}$ en respectant la structure de dépendance réelle des données. **C'est précisément le rôle du VAEAC**, qui apprend à modéliser $p(x_{\bar{S}} \mid x_S, m)$ pour n'importe quel masque m (donc n'importe quelle coalition S).

Vérification des Axiomes de Shapley. L'utilisation de l'espérance conditionnelle $v(S) = \mathbb{E}[f(x) \mid x_S = x_S^*]$ comme fonction caractéristique garantit que les valeurs de Shapley résultantes ϕ_i respectent les propriétés théoriques fondamentales d'une attribution équitable :

1. **Efficacité (Efficiency)** : $\sum_{i=1}^{d} \phi_i = f(x^*) - \mathbb{E}[f(x)]$.

   Démonstration. L'espérance conditionnelle vérifie :

   $$v(N) = \mathbb{E}[f(x) \mid x = x^*] = f(x^*) \quad \text{(conditionnement complet)},$$
   $$v(\emptyset) = \mathbb{E}[f(x)] \quad \text{(aucune information)}.$$

   L'axiome d'efficacité impose $\sum_{i=1}^{d} \phi_i = v(N) - v(\emptyset) = f(x^*) - \mathbb{E}[f(x)]$. □

2. **Symétrie** : Si $\mathbb{E}[f(x) \mid x_S, x_i] = \mathbb{E}[f(x) \mid x_S, x_j]$ pour tout $S \subseteq N \setminus \{i, j\}$, alors $\phi_i = \phi_j$.

   Démonstration. Par hypothèse, $v(S \cup \{i\}) = v(S \cup \{j\})$ pour tout S. Les contributions marginales sont donc identiques, et la formule de Shapley (qui traite i et j symétriquement) donne $\phi_i = \phi_j$. □

3. **Joueur Nul** : Si $\mathbb{E}[f(x) \mid x_S, x_i] = \mathbb{E}[f(x) \mid x_S]$ pour tout $S \subseteq N \setminus \{i\}$, alors $\phi_i = 0$.

   Démonstration. L'hypothèse signifie que x_i n'apporte aucune information conditionnelle : $v(S \cup \{i\}) = v(S)$ pour tout S. Donc toutes les contributions marginales $v(S \cup \{i\}) - v(S) = 0$, impliquant $\phi_i = 0$. □

4. **Linéarité** : Si $f = \alpha f_1 + \beta f_2$, alors $\phi_i(f) = \alpha \phi_i(f_1) + \beta \phi_i(f_2)$.

   Démonstration. Par linéarité de l'espérance conditionnelle :

   $$v(S) = \mathbb{E}[\alpha f_1(x) + \beta f_2(x) \mid x_S = x_S^*] = \alpha \mathbb{E}[f_1(x) \mid x_S = x_S^*] + \beta \mathbb{E}[f_2(x) \mid x_S = x_S^*].$$

   En notant $v_k(S) = \mathbb{E}[f_k(x) \mid x_S = x_S^*]$, on a $v(S) = \alpha v_1(S) + \beta v_2(S)$. La formule de Shapley étant linéaire en v, on obtient directement $\phi_i(f) = \alpha \phi_i(f_1) + \beta \phi_i(f_2)$. □

Remarque 1 (Importance de la qualité d'imputation). Ces axiomes reposent sur l'hypothèse que $v(S)$ est correctement estimée via l'espérance conditionnelle. Or, l'équation (5) montre que $\hat{v}(S) = \frac{1}{K} \sum_{k=1}^{K} f(x_{\bar{S}}^{(k)}, x_S^*)$ avec $x_{\bar{S}}^{(k)} \sim p(x_{\bar{S}} \mid x_S^*)$. Si le VAEAC produit des imputations biaisées — par exemple, en convergeant systématiquement vers la moyenne dans les distributions multimodales (voir Section 6.5) — alors $\hat{v}(S) \neq v(S)$, et les valeurs de Shapley calculées $\hat{\phi}_i$ violent les axiomes en pratique, rendant l'interprétation trompeuse.

Exemple 1.1 (Illustration)**.** Considérons un modèle de prédiction de prix immobilier f avec $x = (\text{surface}, \text{localisation}, \text{nb_pièces})$. Pour expliquer $f(x^*) = 350\text{k}$:
— $v(\emptyset)$: prédiction moyenne quand toutes les variables sont simulées (prix moyen du marché)
— $v(\{\text{surface}\})$: prédiction moyenne quand on fixe la surface à $x^*_{\text{surface}} = 80m^2$, mais qu'on simule localisation et nb_pièces
— $v(\{\text{surface}, \text{localisation}\})$: on fixe aussi la localisation, seul nb_pièces reste incertain
— $v(N)$: tout est fixé, donc $f(x^*) = 350\text{k}$

La valeur de Shapley ϕ_{surface} mesure la contribution moyenne de la surface sur *tous les ordres d'observation possibles* des variables.

2 Auto-Encodeurs Variationnels et Conditionnement

2.1 VAE (Variational Autoencoder)

Le VAE est un modèle génératif probabiliste qui suppose que les données observées x sont générées par des variables latentes non observées z. Contrairement à un auto-encodeur classique qui apprend une projection déterministe, le VAE apprend la distribution conjointe via une optimisation variationnelle.

L'objectif est de maximiser la vraisemblance marginale des données. Cette quantité étant intraitable, on maximise sa borne inférieure (ELBO, *Evidence Lower BOund*) :

$$\log p_\theta(x) \geq \mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p(z)), \tag{6}$$

Définition 2.1 (Architecture VAE)**.** Le modèle repose sur l'articulation de deux réseaux de neurones :

1. **L'Encodeur probabiliste** (Réseau d'inférence) $q_\phi(z|x)$: Il projette la donnée x vers l'espace latent en prédisant les paramètres μ et σ d'une distribution gaussienne : $z \sim \mathcal{N}(\mu(x), \text{diag}(\sigma^2(x)))$.
2. **Le Décodeur probabiliste** (Réseau génératif) $p_\theta(x|z)$: Il reconstruit la distribution de la donnée originale $\hat{x}$ à partir du vecteur latent z.

Pour rendre l'opération d'échantillonnage dérivable (indispensable à la rétropropagation du gradient), on applique l'*astuce de reparamétrisation* :

$$z = \mu + \sigma \odot \epsilon, \quad \text{avec } \epsilon \sim \mathcal{N}(0, I).$$

2.2 VAEAC : VAE with Arbitrary Conditioning

L'architecture VAEAC (*Variational Autoencoder with Arbitrary Conditioning*), introduite par Ivanov et al. (2019), étend le VAE standard pour résoudre le problème d'imputation conditionnelle universelle.

Contrairement aux modèles conditionnels classiques entraînés pour une tâche fixe (ex : prédire y sachant x), le VAEAC est conçu pour estimer la distribution conditionnelle $p(x_{\bar{S}}|x_S)$ pour n'importe quel sous-ensemble de caractéristiques observées $S \subseteq \{1, \ldots, d\}$.

Notations. Pour un vecteur de données $x \in \mathbb{R}^d$:
— $S \subseteq \{1, \ldots, d\}$ désigne l'ensemble des indices observés
— $\bar{S} = \{1, \ldots, d\} \setminus S$ désigne l'ensemble des indices manquants
— $x_S \in \mathbb{R}^{|S|}$ représente le sous-vecteur des variables observées
— $x_{\bar{S}} \in \mathbb{R}^{|\bar{S}|}$ représente le sous-vecteur des variables manquantes
— $m \in \{0, 1\}^d$ est un masque binaire tel que $m_i = 1$ si $i \in S$ et $m_i = 0$ sinon

Le VAEAC prend en entrée le vecteur partiellement observé $(x \odot m, m)$ où $\odot$ désigne le produit de Hadamard (élément par élément).

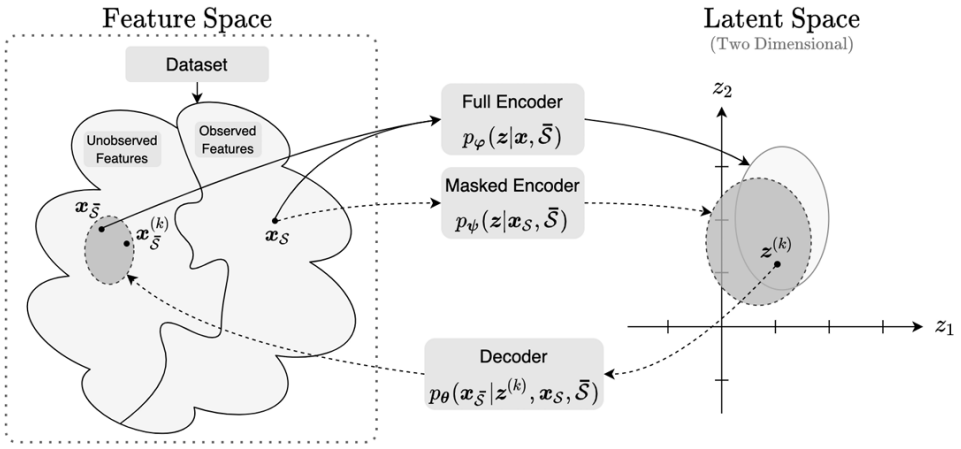


FIGURE 1 – **Processus d'inférence du VAEAC.** Le modèle génère les échantillons manquants $x_{\bar{S}}$ conditionnellement aux observations x_S. Durant l'apprentissage, un "Proposal Network" (encodeur complet) guide l'apprentissage, tandis qu'au déploiement, seul le "Prior Network" (encodeur masqué) est utilisé pour générer les complétions via l'espace latent, tiré du papier original.

2.2.1 Architecture du VAEAC : Deux réseaux principaux

Le VAEAC repose sur une architecture à **deux réseaux d'encodage** et un décodeur, permettant l'apprentissage d'un prior conditionnel.

Définition 2.2 (Composantes du VAEAC)**.** Le modèle articule les trois modules suivants (voir Figure 2) :

1. **Proposal Network (Full Encoder)** $q_\phi(z|x,m)$:
   — Observe la donnée *complète* x et le masque m
   — Utilisé **uniquement durant l'entraînement**
   — Rôle : apprendre une représentation latente optimale en ayant accès à toute l'information
2. **Prior Network (Masked Encoder)** $p_\psi(z|x_S,m)$:
   — N'observe que la partie visible x_S (équivalent à $x \odot m$)
   — Utilisé durant l'**entraînement et le déploiement**
   — Rôle : approximer la distribution latente du Proposal Network sans accès aux données manquantes
3. **Décodeur (Generator)** $p_\theta(x_{\bar{S}}|z,x_S,m)$:
   — Reconstruit les données manquantes à partir du code latent z et des observations x_S
   — Utilisé durant l'**entraînement et le déploiement**

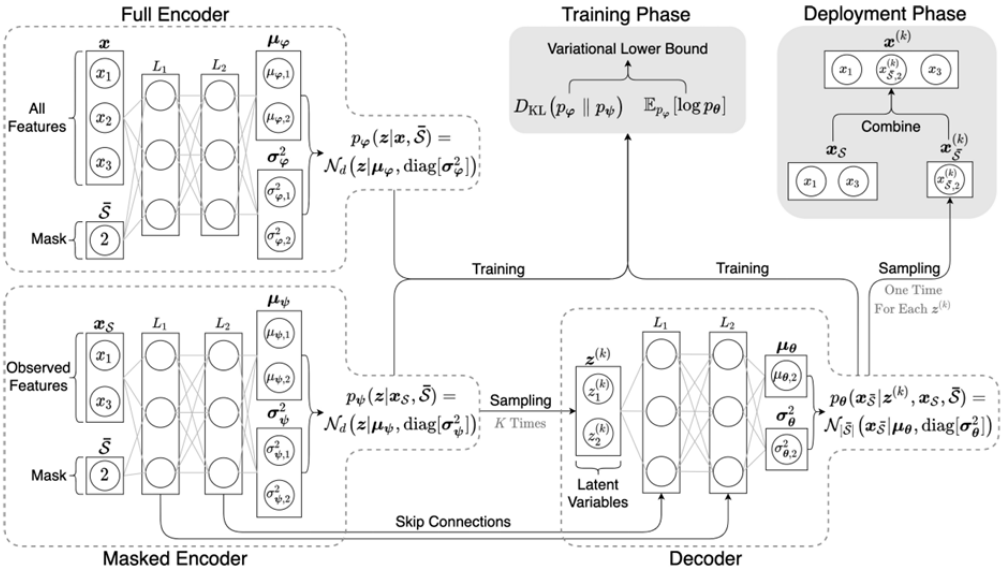


FIGURE 2 – **Schéma détaillé du VAEAC.** Comparaison des flux de données entre l'entraînement (où x complet est connu) et le test (où seul x_S est connu).image tiré du papier original.

2.2.2 Objectif d'optimisation et ELBO conditionnelle

Soit p_{data} la vraie distribution des données. L'objectif est de maximiser la log-vraisemblance conditionnelle des données manquantes, moyennée sur une distribution de masques $p(m)$ (typiquement uniforme ou Bernoulli) :

$$\mathcal{J} = \mathbb{E}_{x \sim p_{\text{data}}} \mathbb{E}_{m \sim p(m)} \left[\log p_{\psi,\theta}(x_{\bar{S}} | x_S, m) \right] \tag{7}$$

Cette quantité étant intraitable, le VAEAC optimise la borne inférieure variationnelle (ELBO) modifiée :

$$\text{ELBO}(x, m) = \mathbb{E}_{q_\phi(z|x,m)} [\log p_\theta(x_{\bar{S}} | z, x_S, m)] - D_{KL}(q_\phi(z|x,m) \| p_\psi(z|x_S, m)) \tag{8}$$

Interprétation des deux termes :

1. **Terme de reconstruction** : $\mathbb{E}_{q_\phi} [\log p_\theta(x_{\bar{S}} | z, x_S, m)]$
   — Pousse le décodeur à générer des complétions plausibles
   — Évalue la qualité de l'imputation
2. **Terme de régularisation (KL)** : $D_{KL}(q_\phi(z|x,m) \| p_\psi(z|x_S, m))$
   — Force la distribution du Masked Encoder (utilisé en test) à se rapprocher de celle du Full Encoder (qui voit la vérité terrain)
   — Évite les imputations hors-distribution
   — Garantit la cohérence entre entraînement et déploiement

Garantie variationnelle. L'ELBO de l'équation (8) est une borne inférieure *exacte*. Pour tout (x, m), on a :

$$\log p_{\theta,\psi}(x_{\bar{S}} \mid x_S, m) = \text{ELBO}(x, m) + D_{KL}\big(q_\phi(z \mid x, m) \,\|\, p_{\theta,\psi}(z \mid x_{\bar{S}}, x_S, m)\big), \tag{9}$$

où le terme KL (toujours positif) mesure le *gap variationnel* entre la distribution d'inférence q_ϕ et la vraie postérieure.

Maximiser l'ELBO revient donc à :

(i) Améliorer la qualité de reconstruction de $x_{\bar{S}}$ via p_θ

(ii) Réduire l'écart entre l'encodeur complet (teacher) et l'encodeur masqué (student)

Remarque 2 (Pondération des masques). Comme l'objectif est moyenné sur $m \sim p(m)$, la convergence est *pondérée* : les conditionnelles associées à des masques rares (ex : $|S| = 1$ ou $|S| = d - 1$) sont moins contraintes et donc potentiellement moins précises pour l'estimation des valeurs de Shapley.

Procédure d'optimisation. En pratique, on minimise la perte négative de l'ELBO moyennée sur les données et les masques :

$$\min_{\theta,\psi,\phi} \mathbb{E}_{x\sim p_{\text{data}}} \mathbb{E}_{m\sim p(m)} \big[-\text{ELBO}(x, m) \big]. \tag{10}$$

Génération conditionnelle (Inférence). Au déploiement, pour imputer $x_{\bar{S}}$ sachant x_S, on marginalise la variable latente z et on approxime par échantillonnage Monte Carlo :

$$p(x_{\bar{S}} \mid x_S, m) \approx \frac{1}{K} \sum_{k=1}^{K} p_\theta(x_{\bar{S}} \mid z^{(k)}, x_S, m) \quad \text{avec} \quad z^{(k)} \sim p_\psi(z \mid x_S, m). \tag{11}$$

Seul le **Masked Encoder** p_ψ est utilisé en test (le Full Encoder q_ϕ n'a plus accès à x complet).

2.3 Rôle Structurel du Masque d'Observation

La conception de l'entrée du réseau mérite une attention particulière. Une observation partielle n'est pas simplement un vecteur x_S, mais un **couple** $(x \odot m, m)$ où :

— $x \odot m$ contient les valeurs observées et des zéros ailleurs

— m est le masque binaire de présence

Cette représentation explicite est cruciale pour lever l'**ambiguïté du zéro** : sans le masque m, le modèle ne peut distinguer :

— Une valeur *observée* égale à zéro : $x_i = 0$ et $m_i = 1$

— Une valeur *manquante* remplacée par zéro : $x_i = 0$ et $m_i = 0$

Par conséquent, les encodeurs du VAEAC ne modélisent pas $q_\phi(z|x)$, mais explicitement :

$$q_\phi(z \mid x, m) \quad \text{(Full Encoder : voit tout)} \tag{12}$$

$$p_\psi(z \mid x \odot m, m) \quad \text{(Masked Encoder : voit uniquement } x_S\text{)} \tag{13}$$

Le masque m agit comme une **variable de contrôle structurelle**. Il modifie la topologie effective de l'entrée, permettant au réseau d'apprendre une famille de distributions latentes indexées par l'ensemble des masques possibles $\mathcal{M} = \{0, 1\}^d$.

Cette dépendance explicite garantit la cohérence de l'inférence : pour un même vecteur de valeurs, deux masques distincts $m_1 \neq m_2$ induisent deux distributions a posteriori $q(z|\cdot, m_1)$ et $q(z|\cdot, m_2)$ potentiellement très différentes, reflétant l'incertitude variable selon le nombre de variables observées.

2.4 Positionnement par rapport aux architectures standards

Le VAEAC peut être formalisé comme une généralisation unificatrice du VAE standard et du Conditional VAE (CVAE). Le tableau 1 compare leurs factorisations respectives.

Modèle	**Factorisation**	**Flexibilité**	**Conditionnement**
VAE	$p(x, z) = p(z)p_\theta(x\|z)$	Rigide	Aucun
CVAE	$p(x, z\|y) = p(z\|y)p_\theta(x\|z, y)$	Semi-flexible	Fixe sur y
VAEAC	$p(x_{\bar{S}}\|x_S, m) = \int p_\theta(\cdot\|z, x_S, m)p_\psi(z\|x_S, m)dz$	Universelle	Arbitraire via m

TABLE 1 – Comparaison des architectures VAE, CVAE et VAEAC

1. VAE Standard. Le modèle suppose que toutes les données sont observées. Il n'y a aucune distinction entre variables d'entrée et de sortie :

$$p(x, z) = p(z)\, p_\theta(x \mid z) \quad \text{et} \quad q_\phi(z \mid x).$$

Limitation : ne permet pas l'inférence partielle sans réentraînement ou approximations coûteuses (comme l'optimisation itérative du latent).

2. Conditional VAE. Le CVAE introduit une variable de conditionnement y (ex : étiquette de classe) qui est *toujours* observée :

$$p(x, z \mid y) = p(z \mid y)\, p_\theta(x \mid z, y) \quad \text{et} \quad q_\phi(z \mid x, y).$$

Limitation : la fixité de y empêche de répondre à des requêtes où une partie de y serait manquante, ou où une partie de x servirait à prédire le reste.

3. VAEAC. Le VAEAC lève la distinction rigide entre condition et cible. Le conditionnement devient dynamiques : n'importe quel sous-ensemble de variables, défini par le masque m, peut servir de conditionnement (x_S) ou de cible ($x_{\bar{S}}$). Le modèle apprend simultanément les 2^d distributions conditionnelles possibles :

$$p(x_{\bar{S}} \mid x_S, m) \approx \int p_\theta(x_{\bar{S}} \mid z, x_S, m)\, p_\psi(z \mid x_S, m)\, dz.$$

Remarque 3 (Interprétation). Le VAEAC peut être vu comme un CVAE dont la variable de conditionnement est **stochastique** (tirée selon $p(m)$) et **combinatoire** (prend 2^d valeurs possibles). Cette généralisation permet une flexibilité totale pour l'imputation conditionnelle.

2.5 Traitement des Données Manquantes : Deux Protocoles Distincts

La présence des données incomplètes dans les applications réelles pose un défi majeur. Nous distinguons ici notre protocole d'évaluation :

2.5.1 Protocole 1 : VAEAC Standard (Marginalisation)

L'architecture VAEAC native traite les données manquantes durant l'apprentissage via un mécanisme probabiliste à deux niveaux :

1. **Conditionnement adaptatif du masque** : La distribution des masques $p(m|x)$ est conditionnée par l'observation elle-même, garantissant que les entrées manquantes (NA) sont systématiquement traitées comme non-observées par le réseau.
2. **Marginalisation de la vraisemblance** : Le calcul de la perte de reconstruction **exclut explicitement** les valeurs manquantes. La log-vraisemblance est évaluée par sommation partielle sur les indices observés uniquement :

$$\log p_\theta(x_{\bar{S}}|z, x_S) = \sum_{i \in \bar{S},\, x_i \neq \text{NA}} \log p_\theta(x_i|z, x_S). \tag{14}$$

Avantage : Le modèle apprend directement sur les données partielles sans biais d'initialisation.

2.6 Métriques d'Évaluation

Pour quantifier la performance des méthodes d'imputation dans le contexte de l'explicabilité (XAI), nous définissons une métrique principale basée sur l'erreur des valeurs de Shapley.

2.6.1 Métriques pour Valeurs de Shapley

Pour quantifier la fidélité de l'estimation des valeurs de Shapley, nous utilisons deux métriques standard d'erreur entre les valeurs estimées $\hat{\phi}_i$ et les valeurs de référence ϕ_i^{true} :

Définition 2.3 (Mean Absolute Error (MAE))**.** L'erreur absolue moyenne mesure l'écart moyen en valeur absolue :

$$\text{MAE} = \frac{1}{N \cdot d} \sum_{n=1}^{N} \sum_{i=1}^{d} \left| \phi_i^{\text{true}}(x^{(n)}) - \hat{\phi}_i(x^{(n)}) \right|,$$

où N est le nombre d'instances évaluées et d le nombre de features.

Définition 2.4 (Mean Squared Error (MSE))**.** L'erreur quadratique moyenne pénalise davantage les grandes déviations :

$$\text{MSE} = \frac{1}{N \cdot d} \sum_{n=1}^{N} \sum_{i=1}^{d} \left(\phi_i^{\text{true}}(x^{(n)}) - \hat{\phi}_i(x^{(n)}) \right)^2.$$

Remarque 4. La MAE est robuste aux valeurs aberrantes et préserve l'échelle originale des contributions. La MSE, plus sensible aux erreurs importantes, est particulièrement utile pour détecter les cas où le modèle d'imputation échoue à capturer la structure de dépendance (ex : distribution multimodale lissée vers la moyenne).

2.6.2 Métriques de Reconstruction Auxiliaires

Pour l'analyse de sensibilité et la validation intermédiaire, nous utilisons également :

RMSE : Mesure la qualité de reconstruction vis à vis des vraies données.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{i,j} - \hat{x}_{i,j})^2} \tag{15}$$

Ecart-type La stabilité de la méthode est quantifiée par l'écart-type (σ) calculé sur N itérations indépendantes :

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (\text{RMSE}_k - \overline{\text{RMSE}})^2} \tag{16}$$

Une faible RMSE indique une grande précision, tandis qu'un faible σ indique une robustesse aux conditions initiales (stabilité algorithmique).

Divergence de Kullback-Leibler (KL) : Mesure la distance entre la distribution reconstruite et la vraie distribution

$$D_{KL}(p_{\text{true}} \| p_{\text{rec}}) = \int p_{\text{true}}(x) \log \frac{p_{\text{true}}(x)}{p_{\text{rec}}(x)} dx \tag{17}$$

Coefficient de détermination (R^2) : Mesure la qualité de prédiction

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{18}$$

2.7 Fonction de Coût et Régularisation

L'entraînement maximise l'ELBO conditionnelle. La fonction de perte totale $\mathcal{L}$ inclut un terme de régularisation Ω sur les paramètres du prior pour stabiliser le début de l'apprentissage :

$$\mathcal{L} = -\mathbb{E}_{q_\phi}[\log p_\theta(x|z)] + D_{KL}(q_\phi(z|x)||p_\psi(z|x_{obs}, b)) + \lambda\Omega(\lambda << 1), \tag{19}$$

3 Analyse de la Qualité de Reconstruction

3.1 Résultats Expérimentaux

Le tableau ci-dessous synthétise les performances de reconstruction (Moyenne ± Écart-type) pour cinq méthodes d'imputation sur six datasets distincts. La meilleure performance par dataset est indiquée en gras.

TABLE 2 – Comparaison de la RMSE de reconstruction des valeurs de Shapley (Moyenne ± Std Dev) sur différents datasets.

Dataset	Mean	Zero	KNN	MissForest	VAEAC
ABALONE	**0.7024 ± 0.0618**	0.7103 ± 0.0333	1.0832 ± 0.1390	1.1265 ± 0.1168	0.7463 ± 0.0397
CALIFORNIA	0.9695 ± 0.0297	**0.8882 ± 0.0448**	0.9491 ± 0.0152	0.9113 ± 0.0497	0.9407 ± 0.0035
DIABETES	0.8751 ± 0.0519	0.8833 ± 0.0444	0.8594 ± 0.0310	**0.8445 ± 0.0159**	0.8902 ± 0.0131
IID	1.0035 ± 0.0082	1.0093 ± 0.0051	1.0087 ± 0.0196	1.0036 ± 0.0034	**0.9948 ± 0.0037**
GMM	0.8810 ± 0.0322	0.8471 ± 0.0138	0.8493 ± 0.0342	**0.8312 ± 0.0055**	0.8502 ± 0.0120
COPULA	**0.6230 ± 0.0477**	0.7388 ± 0.0231	0.6482 ± 0.0320	0.6581 ± 0.0308	0.6922 ± 0.0751

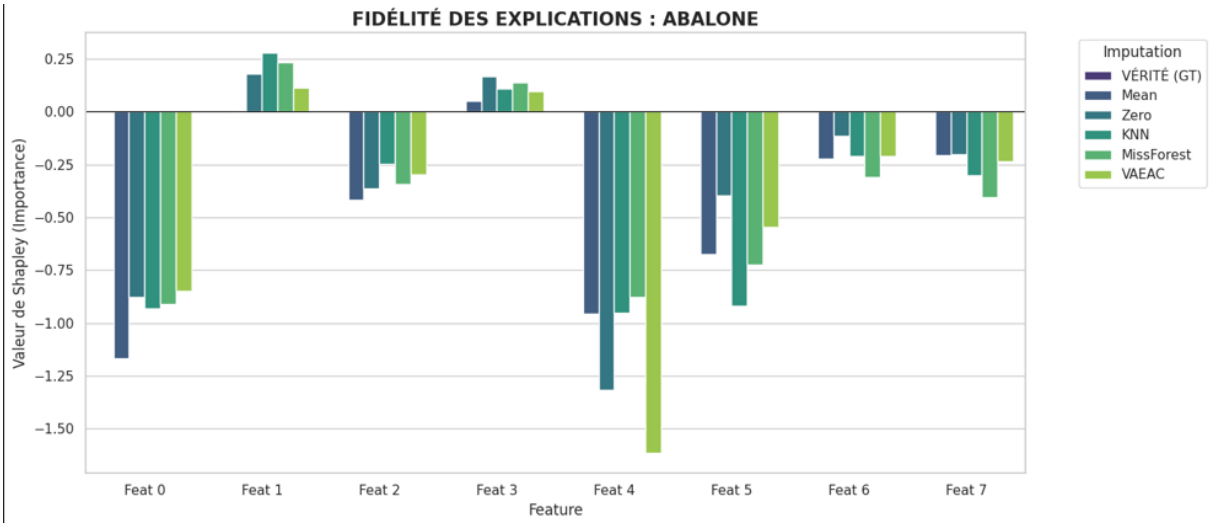


FIGURE 3 – Évaluation de la fidélité des valeurs de Shapley (Moyenne ± Écart-type) selon la méthode d'imputation sur le dataset *Abalone*. Les résultats sont agrégés sur 3 itérations (*runs*) indépendantes.

3.2 Analyse de la Fidélité Explicative

L'impact du choix de la méthode d'imputation sur l'explication locale par valeurs de Shapley est illustré dans la Figure 3. On observe que :

— **Stabilité :** Les méthodes **MissForest** (en gris) et **VAEAC** (en cyan) présentent des barres d'erreur ($\pm\sigma$) plus marquées sur les caractéristiques à forte contribution (ex : F3, F5), reflétant la variance inhérente à ces modèles stochastiques.

— **Biais de Neutralité :** Les méthodes **Mean** et **Zero** tendent à sous-estimer systématiquement l'importance de certaines variables clés, en ancrant les valeurs manquantes vers des zones de faible information.

— **Convergence :** Malgré l'incertitude liée à la reconstruction, la hiérarchie des contributions (le signe et le rang des features) reste globalement cohérente entre MissForest et VAEAC, validant la robustesse du couplage entre modèles génératifs et théorie des jeux.

TABLE 3 – Détail des performances (RMSE) selon la dimension du masque (k).

		RMSE par Méthode		
Dataset	**Nb. Features (k)**	**Mean**	**MissForest**	**VAEAC**
COPULA	1	**0.525**	0.748	0.636
	2	0.706	0.796	**0.699**
	4	**0.623**	0.684	0.625
	8	**0.652**	0.680	0.711
GMM	1	**0.653**	0.677	0.666
	2	**0.649**	0.697	0.662
	4	**0.646**	0.689	0.680
	8	0.721	**0.675**	0.727
IID	1	1.154	1.120	**1.106**
	2	1.119	1.148	**1.117**
	4	1.118	**1.097**	1.114
	8	1.091	**1.082**	1.086

3.3 Impact de la Dimension du Masque

L'analyse conjointe de la Figure 4 et du Tableau 3 met en évidence trois comportements distincts face à l'augmentation de l'incertitude locale (k) :

Dataset IID (variables indépendantes). Les scores RMSE sont systématiquement élevés (> 1.0) et montrent une faible sensibilité à k. Ce comportement est cohérent avec l'absence de structure de dépendance : puisque les variables sont indépendantes par construction, observer x_S n'apporte aucune information sur $x_{\bar{S}}$. Les trois méthodes convergent vers des performances similaires, le VAEAC capturant correctement la distribution marginale sans halluciner de corrélations inexistantes. L'absence de dégradation avec k croissant confirme que le problème est intrinsèquement non-informatif.

Dataset GMM (structure multimodale). On observe une dégradation modérée avec l'augmentation de k, particulièrement visible pour les méthodes paramétriques. À $k = 8$ (occultation forte), MissForest (RMSE = 0.675) surpasse significativement Mean (0.721) et VAEAC (0.727). Cette robustesse s'explique par la capacité des arbres de décision à partitionner l'espace de manière non-paramétrique, préservant la séparation entre modes même lorsque l'information conditionnelle devient rare. Le VAEAC, contraint par son prior gaussien, tend à lisser les transitions entre modes (voir Section 6.5).

Dataset Copula (dépendances non-linéaires complexes). Ce cas révèle le paradoxe de la métrique RMSE. L'imputation par la moyenne affiche les meilleurs scores (0.525 à $k = 1$, 0.623 à $k = 4$), mais cette performance est trompeuse : en ancrant systématiquement les prédictions sur $\mathbb{E}[x_{\bar{S}}|x_S]$, elle minimise mécaniquement l'erreur quadratique au détriment de la variance conditionnelle. MissForest et VAEAC, qui tentent de préserver la structure distributionnelle (multimodalité, queues de distribution), exhibent des RMSE légèrement supérieures mais génèrent des échantillons plus fidèles à $p(x_{\bar{S}}|x_S)$. Cette observation justifie l'introduction de métriques complémentaires comme la divergence KL pour évaluer la qualité distributionnelle au-delà de l'erreur ponctuelle.

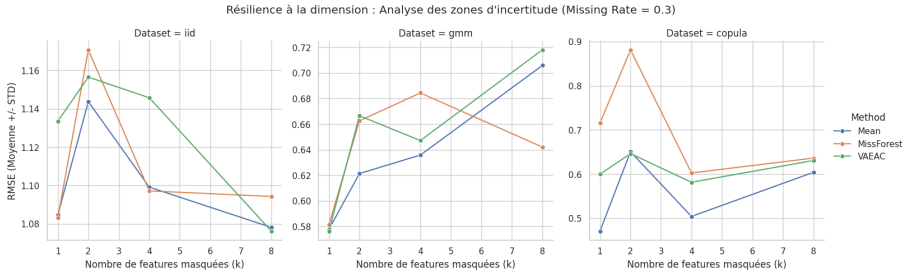


FIGURE 4 – **Analyse des zones d'incertitude :** Évolution de l'erreur de reconstruction (RMSE) en fonction du nombre de caractéristiques masquées ($k \in \{1, 2, 4, 8\}$) pour les datasets IID, GMM et Copula. Le taux de données manquantes global est fixé à 30%.

4 Analyse de Sensibilité : Impact de ϵ

4.1 Visualisation des Performances

L'évolution des métriques en fonction du paramètre de bruit $\epsilon \in \{0.00, 0.50, 0.99\}$ est illustrée dans la Figure 5.

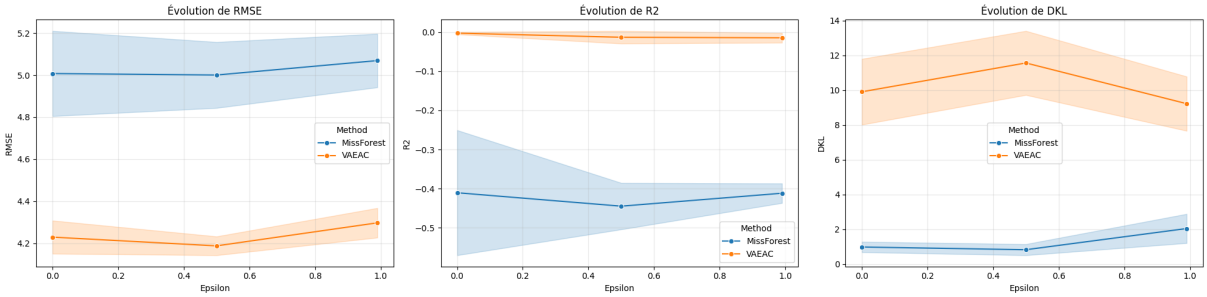


FIGURE 5 – **Dynamique de l'Imputation :** Comparaison de l'évolution de la RMSE, du R^2 et de la Divergence KL pour MissForest et VAEAC. Les zones ombrées représentent l'intervalle de confiance (Std Dev).

Le tableau ci-dessous détaille les résultats numériques (Moyenne ± Écart-type) pour les trois niveaux de difficulté définis par ϵ.

TABLE 4 – Performance comparée selon ϵ (Moyenne ± Std Dev sur 3 runs).

Epsilon (ϵ)	**Méthode**	**RMSE**	R^2	**DKL**
0.00	MissForest	5.0074 ± 0.2027	−0.4103 ± 0.1597	**0.9735 ± 0.2998**
	VAEAC	**4.2290 ± 0.0788**	**−0.0026 ± 0.0030**	9.9047 ± 1.8920
0.50	MissForest	5.0003 ± 0.1568	−0.4446 ± 0.0593	**0.8175 ± 0.3199**
	VAEAC	**4.1875 ± 0.0448**	**−0.0132 ± 0.0162**	11.5641 ± 1.8427
0.99	MissForest	5.0690 ± 0.1274	−0.4115 ± 0.0250	**2.0368 ± 0.8392**
	VAEAC	**4.2970 ± 0.0703**	**−0.0145 ± 0.0124**	9.2211 ± 1.5621

Analyse des résultats. **Domination du VAEAC sur les métriques ponctuelles.** Quelle que soit la structure de corrélation ($\epsilon \in \{0.00, 0.50, 0.99\}$), le VAEAC surpasse systématiquement MissForest en termes d'erreur quadratique (RMSE ≈ 4.2 contre ≈ 5.0, soit une réduction de ∼16%). De plus, son score R^2 reste proche de zéro (−0.0026 à −0.0145), indiquant que le modèle prédit une valeur proche de la

moyenne globale (qui vaut 0 par symétrie du mélange). En revanche, MissForest s'effondre avec des R^2 fortement négatifs (≈ -0.41), signalant que ses prédictions dévient systématiquement de l'espérance conditionnelle. Paradoxalement, ce résultat pourrait refléter une qualité supérieure : en tentant de capturer l'un des deux modes (± 5) plutôt que leur moyenne (0), MissForest génère des échantillons plus réalistes mais qui maximisent mécaniquement la MSE.

la divergence KL : signature du lissage. Inversement, MissForest présente une divergence de Kullback-Leibler nettement inférieure (DKL $\approx$ 0.8–2.0) comparée au VAEAC (DKL $\approx$ 9.2–11.6, soit un facteur $\times 5$). Cette observation confirme le diagnostic de lissage établi en Section 6.5. Le VAEAC, minimisant la perte L^2 avec un décodeur gaussien, converge vers l'espérance conditionnelle. Dans le cas où $\epsilon = 0$ (variables indépendantes, $\Sigma_{S\bar{S}} = 0$), l'analyse de la Section 6.5.3 démontre que $\mathbb{E}[x_{\bar{S}}|x_S] \approx 0$ dans les régions d'ambiguïté, ce qui explique :

— La faible RMSE : prédire la moyenne minimise mécaniquement l'erreur quadratique,

— La forte DKL : générer systématiquement des valeurs proches de 0 (creux de densité entre les modes ± 5) produit une distribution reconstruite concentrée dans une région de probabilité quasi-nulle sous la vraie distribution $p(x)$.

MissForest, grâce à son partitionnement non-paramétrique, génère des échantillons appartenant majoritairement aux modes réels (± 5), préservant ainsi la forme bimodale (faible DKL) au prix d'une variance élevée dans les prédictions ponctuelles (forte RMSE).

5 Stress-Test de Robustesse : Censure MNAR

5.1 Protocole de Censure et Biais

Contrairement à l'hypothèse MCAR (*Missing Completely At Random*) où l'absence d'information est aléatoire, le scénario MNAR (*Missing Not At Random*) introduit une dépendance directe entre la valeur d'une variable et sa probabilité d'être manquante. Ce mécanisme simule des contraintes physiques réelles, telles qu'un capteur saturé ou la dégradation sélective d'échantillons anciens.

Pour évaluer la capacité du modèle à extrapoler hors de sa zone d'entraînement (domaine observé), nous avons appliqué une censure déterministe sur la variable `WholeWeight`. La probabilité de masquage S pour une observation $X_{i,j}$ est régie par une fonction logistique :

$$P(S_{i,j} = 1|X_{i,j}) = \frac{1}{1 + \exp(-X_{i,j})} \tag{20}$$

Concrètement, nous avons censuré systématiquement les individus situés au-delà du **80ème percentile**. Le défi pour le VAEAC est donc de reconstruire la queue de distribution des individus les plus lourds sans jamais les avoir observés complets durant l'apprentissage.

5.2 Résultats Quantitatifs

Le tableau ci-dessous compare les performances du modèle sous un masquage aléatoire (MCAR, 20%) face à cette censure dirigée (MNAR, Top 20%).

Table 5 – Impact du mécanisme de censure sur la reconstruction et l'explicabilité (Moyenne ± Écart-type).

Scénario	RMSE	R^2	DKL	MAE Shapley	MSE Shapley
MCAR	3.5984 ± 0.04	0.3664 ± 0.03	4.4013 ± 2.35	5.1518 ± 0.23	44.7866 ± 1.65
MNAR	7.1618 ± 0.33	**−127.129 ± 11.87**	1.6151 ± 0.75	**5.1186 ± 0.04**	47.4914 ± 1.97

5.3 Analyse : L'Échec de l'Extrapolation face à la Stabilité Explicative

L'analyse des résultats révèle une dichotomie frappante entre la qualité de la prédiction brute et la cohérence de l'explication fournie par le modèle.

Effondrement de la Reconstruction. Le passage au scénario MNAR provoque une dégradation critique des métriques de reconstruction. La RMSE double quasiment, passant de 3.60 à 7.16. Plus alarmant, le score R^2 plonge à une valeur abyssale de -127.13. Statistiquement, cela indique que les prédictions du modèle sont nettement moins précises qu'une simple imputation par la moyenne globale. Cet échec démontre l'incapacité du VAEAC "pur" à extrapoler : privé des exemples de poids lourds, le modèle ancre ses prédictions sur la moyenne des données observées (les individus légers et moyens), créant un biais massif lors de l'évaluation sur les queues de distribution réelles.

Le Paradoxe de la Divergence KL. Il est intéressant de noter que la Divergence de Kullback-Leibler (DKL) est paradoxalement plus faible en MNAR (1.62) qu'en MCAR (4.40). Ce résultat est trompeur : il ne signale pas une meilleure adéquation, mais plutôt que le modèle génère une distribution très confiante (faible entropie), mais totalement décalée par rapport à la vérité terrain. C'est le symptôme d'un modèle "précis mais faux", qui a appris une distribution tronquée sans capturer l'incertitude liée à la censure.

Robustesse de l'Explicabilité. Malgré l'effondrement des valeurs prédites, la fidélité des valeurs de Shapley (MAE Shapley) reste remarquablement stable, passant de 5.15 en MCAR à 5.12 en MNAR. Ce résultat suggère que si le modèle échoue à quantifier l'intensité exacte des valeurs manquantes, il préserve néanmoins la structure des relations entre variables. Le VAEAC comprend toujours que le poids est une variable déterminante, même s'il en sous-estime la magnitude. La structure logique du modèle survit donc au biais d'échantillonnage, là où la précision numérique échoue.

6 Analyse Théorique : Le problème du Lissage

6.1 Cadre simplifié pour l'analyse

Pour isoler et comprendre le phénomène fondamental de lissage, nous analysons d'abord un cas théorique simplifié : un VAEAC avec un décodeur gaussien à variance unitaire fixe ($\sigma^2 = 1$). Cette simplification permet de :

— **Isoler l'effet de la minimisation** L^2 : En fixant $\sigma^2 = 1$, la maximisation de la vraisemblance se réduit à la minimisation de l'erreur quadratique.
— **Obtenir des résultats analytiques clairs** : La solution optimale devient l'espérance conditionnelle, dont les propriétés sont bien comprises.
— **Comprendre le comportement limite** : Ce cas représente la limite où les régularisations sur la variance sont infiniment fortes.

Lien avec notre implémentation : Dans nos expériences, l'initialisation à zéro des poids de `dec_logvar` (donnant initialement $\sigma^2 = 1$) et les faibles régularisations ($\sigma_\mu = \sigma_\sigma = 10^4$) font que le modèle se comporte approximativement comme ce cas simplifié, validant ainsi la pertinence de notre analyse.

Après avoir établi ce résultat fondamental, nous discuterons dans la section 4.5 comment le relâchement de ces hypothèses (variance apprise, régularisation modale) affecte le phénomène.

Nous formalisons ici le cas spécifique d'une distribution bimodale symétrique, identifiée comme source d'erreurs fréquentes pour les modèles génératifs continus minimisant une perte L_2.

Soit une variable aléatoire X suivant un mélange de deux Gaussiennes de poids égaux, d'espérances opposées μ et $-\mu$ (avec $\mu \gg \sigma$), et de même variance σ^2. La densité s'écrit :

$$p(x) = \frac{1}{2}\mathcal{N}(x; -\mu, \sigma^2) + \frac{1}{2}\mathcal{N}(x; \mu, \sigma^2) \tag{21}$$

Le VAEAC repose sur l'optimisation d'une borne inférieure (ELBO) incluant un terme de reconstruction, généralement l'Erreur Quadratique Moyenne (MSE) pour des données continues. Pour un point latent ambigu z (situé à la frontière de décision dans l'espace latent), le décodeur cherche une valeur $\hat{x}$ minimisant l'espérance de l'erreur :

$$\hat{x}^* = \arg\min_{\hat{x}} \mathbb{E}_{x\sim p(x)} \left[\|x - \hat{x}\|^2\right]$$

La solution analytique de ce problème d'optimisation est l'espérance mathématique de la distribution :

$$\hat{x}^* = \mathbb{E}[X] = \frac{1}{2}(-\mu) + \frac{1}{2}(\mu) = 0$$

6.1.1 Préservation de la symétrie et de la multimodalité par conditionnement

En dimension 1 le résultat est immédiat, le masquage est total (tout ou rien). Lorsqu'on masque l'unique variable, le modèle ne dispose d'aucune information conditionnelle. Il est donc contraint d'apprendre la distribution marginale $P(x)$ pure du jeu de données (le "profil" global des données) au lieu d'une distribution conditionnelle $P(x|y)$

6.1.2 Vecteurs gaussiens quelconques

Soit $X = (X_S, X_{\bar{S}}) \in \mathbb{R}^d$ un vecteur aléatoire suivant un mélange gaussien symétrique centré à deux composantes :

$$p(x) = \frac{1}{2}\mathcal{N}(x; \mu, \Sigma) + \frac{1}{2}\mathcal{N}(x; -\mu, \Sigma), \tag{22}$$

où $\mu \in \mathbb{R}^d$ et Σ est une matrice de covariance définie positive.

On partitionne le vecteur moyen et la matrice de covariance selon les indices observés S et manquants $\bar{S}$:

$$\mu = \begin{pmatrix} \mu_S \\ \mu_{\bar{S}} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{S\bar{S}} \\ \Sigma_{\bar{S}S} & \Sigma_{\bar{S}\bar{S}} \end{pmatrix}.$$

Distribution conditionnelle. La loi conditionnelle d'un mélange gaussien est elle-même un mélange gaussien :

$$p(x_{\bar{S}} \mid x_S) = \sum_{k=1}^{2} w_k(x_S)\, \mathcal{N}(x_{\bar{S}}; \hat{\mu}_k(x_S), \hat{\Sigma}), \tag{23}$$

où la covariance conditionnelle est commune aux deux composantes :

$$\hat{\Sigma} = \Sigma_{\bar{S}\bar{S}} - \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\Sigma_{S\bar{S}}.$$

Moyennes conditionnelles. Pour la composante centrée en $+\mu$:

$$\hat{\mu}_1(x_S) = \mu_{\bar{S}} + \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}(x_S - \mu_S). \tag{24}$$

Pour la composante centrée en $-\mu$:

$$\hat{\mu}_2(x_S) = -\mu_{\bar{S}} + \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}(x_S + \mu_S). \tag{25}$$

Poids a posteriori. Les poids conditionnels s'écrivent :

$$w_k(x_S) = \frac{\frac{1}{2}\mathcal{N}(x_S; \pm\mu_S, \Sigma_{SS})}{\frac{1}{2}\mathcal{N}(x_S; \mu_S, \Sigma_{SS}) + \frac{1}{2}\mathcal{N}(x_S; -\mu_S, \Sigma_{SS})}.$$

Cas symétrique critique : $x_S = 0$. Lorsque l'observation se situe exactement au point de symétrie, on obtient :

$$\hat{\mu}_1(0) = \mu_{\bar{S}} - \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\mu_S, \tag{26}$$

$$\hat{\mu}_2(0) = -\mu_{\bar{S}} + \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\mu_S = -\hat{\mu}_1(0), \tag{27}$$

$$w_1(0) = w_2(0) = \frac{1}{2}. \tag{28}$$

Paradoxe apparent : multimodalité préservée mais lissage inévitable. La distribution conditionnelle $p(x_{\bar{S}} \mid x_S = 0)$ est effectivement un mélange gaussien symétrique bimodal. Cependant, l'**espérance conditionnelle** – qui est la quantité minimisée par un décodeur gaussien avec perte L^2 – vaut :

$$\mathbb{E}[x_{\bar{S}} \mid x_S = 0] = w_1(0)\hat{\mu}_1(0) + w_2(0)\hat{\mu}_2(0) = \frac{1}{2}\hat{\mu}_1(0) + \frac{1}{2}(-\hat{\mu}_1(0)) = 0.$$

Ce résultat est fondamental : même lorsque la distribution conditionnelle est parfaitement bimodale, un modèle qui minimise l'erreur quadratique (ou maximise la vraisemblance gaussienne) produira systématiquement la moyenne des deux modes, c'est-à-dire une valeur située dans le *creux* de la densité.

6.2 Preuve : La Minimisation L^2 induit le Lissage Conditionnel

Dans le cadre simplifié d'un décodeur gaussien à variance unitaire fixe ($\sigma^2 = 1$), la maximisation de la vraisemblance se réduit à la minimisation de l'espérance de l'erreur quadratique sur les données manquantes. Pour une observation partielle x_S donnée, le décodeur cherche une valeur $\hat{x}_{\bar{S}}$ minimisant le risque conditionnel :

$$J(\hat{x}_{\bar{S}}) = \mathbb{E}_{x_{\bar{S}} \sim p(\cdot|x_S)} \left[\|x_{\bar{S}} - \hat{x}_{\bar{S}}\|^2 \right] \qquad (34)$$

Le minimiseur de cette fonction est l'espérance conditionnelle de la distribution. Formellement, en annulant le gradient par rapport à l'imputation $\hat{x}_{\bar{S}}$:

$$\begin{aligned} \nabla_{\hat{x}_{\bar{S}}} J(\hat{x}_{\bar{S}}) &= \nabla_{\hat{x}_{\bar{S}}} \int \|x_{\bar{S}} - \hat{x}_{\bar{S}}\|^2 p(x_{\bar{S}}|x_S) dx_{\bar{S}} \\ &= \int -2(x_{\bar{S}} - \hat{x}_{\bar{S}}) p(x_{\bar{S}}|x_S) dx_{\bar{S}} = 0 \qquad (35) \\ &\Rightarrow \hat{x}^*_{\bar{S}} = \mathbb{E}[x_{\bar{S}}|x_S] \end{aligned}$$

Impact de la covariance nulle dans nos simulations. Dans nos expériences sur données synthétiques, nous avons modélisé des mélanges de gaussiennes avec des matrices de covariance **diagonales** (où éléments hors diagonales presque nuls, i.e peu corrélés). Cela implique que la covariance croisée entre les variables observées et manquantes est **nulle** ($\Sigma_{S\bar{S}} = 0$).

Cette propriété structurelle a deux conséquences majeures :

1. Les variables sont conditionnellement indépendantes sachant le mode latent. L'observation de x_S n'apporte aucune information sur la position spatiale des modes dans l'espace $\bar{S}$ (pas de correction linéaire possible).
2. L'inférence repose uniquement sur les probabilités d'appartenance aux modes (les poids a posteriori $w_k(x_S)$).

Dans les régions d'ambiguïté (quand x_S se trouve entre les modes), les poids des deux modes s'équilibrent ($w_1 \approx w_2 \approx 0.5$). En l'absence de corrélation pour briser la symétrie, la loi conditionnelle reste un mélange parfaitement symétrique. Le minimiseur L^2 converge alors inévitablement vers la moyenne de ce mélange :

$$\hat{x}^*_{\bar{S}} = \mathbb{E}[x_{\bar{S}}|x_S] \approx 0 \qquad (36)$$

Ce résultat démontre que l'échec du VAEAC dans nos simulations n'est pas un défaut d'apprentissage, mais une conséquence mathématique directe du coix du prior gaussien sur une distribution dont la symétrie est protégée par une covariance nulle.

6.3 Peut-on éviter le flou ?

Le phénomène de flou observé n'est pas dû à une mauvaise optimisation ni à une capacité insuffisante du réseau, mais constitue une conséquence structurelle de la minimisation d'une perte quadratique sur une loi conditionnelle multimodale. Il est donc naturel de se demander s'il existe des stratégies permettant de préserver la multimodalité des solutions.

1. Modéliser explicitement une loi conditionnelle multimodale. Une solution plus principielle consiste à enrichir la famille de distributions conditionnelles $p_\theta(x_{\bar{S}} \mid x_S, z)$, par exemple en :
— utilisant un décodeur à mélange de gaussiennes,
— introduisant des variables latentes discrètes capturant le choix du mode,
— combinant le modèle avec des normalizing flows conditionnels.
Ces approches permettent au modèle de représenter plusieurs solutions plausibles au lieu d'en moyenner une seule.

2. Meilleure Imputation. Une meilleure imputation permet de préserver au mieux la loi , minimisant ainsi le phénomène de flou, au prix d'un prix de calcul parfois énorme.

3. Guidage par information supplémentaire. Le flou apparaît principalement dans les régions d'ambiguïté, lorsque l'observation x_S ne contient pas suffisamment d'information pour lever la symétrie entre les modes. Introduire des signaux supplémentaires permet de désambiguïser le conditionnement et de guider le modèle vers une solution cohérente.

6.3.1 Succès de MissForest : Partitionnement vs Continuité

Contrairement aux réseaux de neurones qui sont des approximateurs universels continus, un arbre de décision partitionne l'espace de manière discrète via des fonctions indicatrices $\mathbb{I}$. Cette structure algorithmique confère à MissForest une capacité intrinsèque à préserver les discontinuités de la densité, là où le VAEAC, par sa nature continue et différentiable, tend structurellement au lissage.

Pour imputer une variable cible X_j conditionnellement aux autres variables X_{-j}, chaque arbre de la forêt effectue des coupures nettes (*Hard Splits*) basées sur des seuils optimaux τ. Dans un contexte multimodal bien séparé, ces seuils permettent de discriminer efficacement entre les modes :

$$\hat{f}_t(x_{-j}) = \begin{cases} \hat{\mu}_L \approx -\mu & \text{si } x_{-j} < \tau_t \\ \hat{\mu}_R \approx \mu & \text{si } x_{-j} \geq \tau_t \end{cases} \tag{29}$$

Comportement à la frontière de décision. Dans la région d'ambiguïté (où $x_{-j} \approx 0$), chaque arbre individuel prend une décision binaire basée sur son propre seuil τ_t. La prédiction agrégée de la forêt résulte alors d'un vote majoritaire (ou d'une moyenne de valeurs discrètes) entre les deux modes. Contrairement au VAEAC qui produit systématiquement une valeur intermédiaire (la moyenne $\mathbb{E}[X]$) pour minimiser l'erreur globale, MissForest génère des imputations qui appartiennent majoritairement à l'un ou l'autre des modes, préservant ainsi la structure bimodale de la distribution.

Limite théorique et réalité pratique. Il convient de noter une nuance théorique : si les seuils τ_t étaient distribués de manière parfaitement symétrique autour de 0 et si la forêt était de taille infinie, la prédiction moyenne à $x_{-j} = 0$ convergerait asymptotiquement vers 0 par la loi des grands nombres. Cependant, ce cas limite théorique ne se réalise pas en pratique pour plusieurs raisons :

1. L'entraînement sur des sous-ensembles *bootstrap* induit une variance dans les seuils τ_t, brisant la symétrie parfaite.
2. La profondeur limitée des arbres empêche un partitionnement infiniment fin de l'espace.

Empiriquement, MissForest démontre une capacité supérieure à capturer la multimodalité comparé au VAEAC. Cette performance s'explique par sa nature discrète et non-paramétrique, qui évite le lissage inhérent aux modèles continus minimisant une perte L^2. Néanmoins, cette préservation n'est pas une garantie mathématique absolue dans les régions de forte ambiguïté où l'information discriminante est absente.

6.4 Proposition de Solution : Régularisation par Ancrage Modal

Nous nous replaçons dans le cas critique d'une distribution bimodale symétrique, identifiée en section 6.1 comme la cause de l'échec de la reconstruction (convergence vers la moyenne nulle). Soit X une variable aléatoire suivant un mélange de deux Gaussiennes d'espérances μ et $-\mu$ ($\mu \gg \sigma$).

Plutôt que de modifier l'architecture du prior qui complexifierait l'inférence, nous proposons d'intervenir directement sur le problème d'optimisation en introduisant une pénalité géométrique.

6.4.1 Formulation Mathématique

L'idée est d'ajouter un terme de régularisation Ω qui pénalise les prédictions éloignées des modes de densité connus. Pour un cas bimodal avec des centres a et b, nous définissons la pénalité géométrique :

$$\Omega(y) = \min\left((y-b)^2, (y-a)^2\right)$$

Soit $\mathcal{L}_{rec}(y)$ la fonction de perte de reconstruction standard (par exemple la Log-Vraisemblance Négative gaussienne). Le problème d'optimisation global devient :

$$\hat{x}^* = \text{argmin}\, y \left(\mathcal{L}rec(y) + \lambda\Omega(y)\right), \quad \lambda \geq 0 \tag{30}$$

Notons y_{lisse} le minimiseur de la perte standard non régularisée. Comme démontré précédemment, ce minimiseur correspond à l'espérance conditionnelle , responsable du phénomène de lissage :

$$y_{\text{lisse}} = \text{argmin}_y\, \mathcal{L}_{rec}(y) = \mathbb{E}[x_{\bar{S}}|x_S]$$

En supposant que la solution se situe dans le bassin d'attraction du mode b (i.e., $\Omega(y) = (y-b)^2$) et que la perte de reconstruction est localement quadratique autour de son minimum (on se place dans les hypotheses précédentes), la condition de premier ordre s'écrit :

$$\nabla\mathcal{L}total(y) = 0 \iff \nabla\mathcal{L}rec(y) + \lambda\nabla\Omega(y) = 0$$

$$\iff 2(y - y_{\text{lisse}}) + 2\lambda(y-b) = 0 \iff (1+\lambda)y = y_{\text{lisse}} + \lambda b$$

D'où la solution analytique explicite, exprimée comme une interpolation :

$$y = \frac{y_{\text{lisse}} + \lambda b}{1+\lambda}. \tag{31}$$

Remarque 5. Cette formulation offre une interpolation contrôlable :
— Si $\lambda \to 0$, on retrouve la solution du VAEAC classique ($y = \mathbb{E}[x_{\bar{S}}]$), qui échoue en visant la moyenne.
— Si $\lambda \to \infty$, la prédiction est projetée "durement" sur le mode le plus proche b.

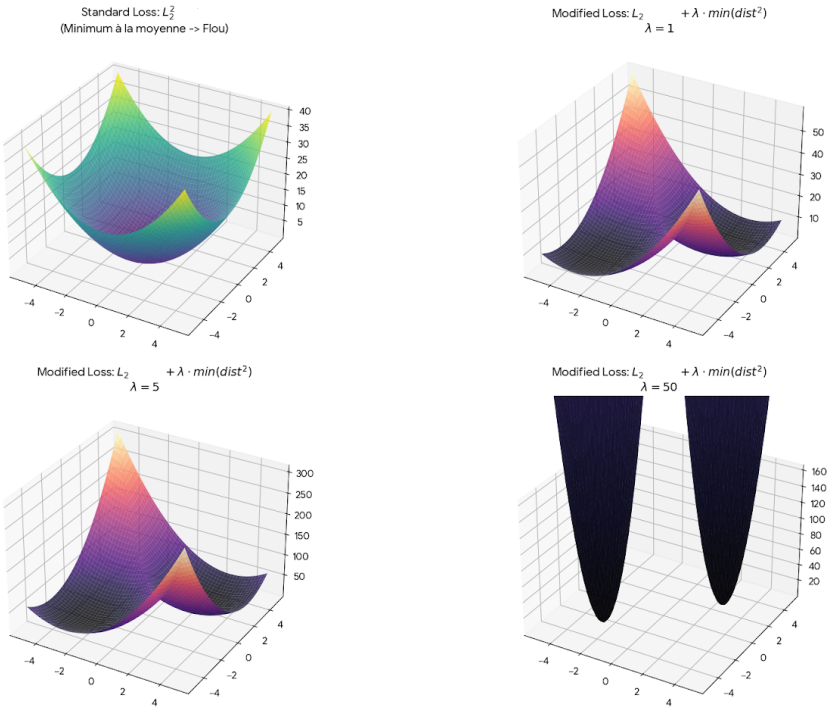


FIGURE 6 – **Changement d'énergie du modèle.** Loss standard (L_2^2) et loss modifiée $L_2^2 + \lambda$ min(dist^2).

Le choix du lambda change le problème d'optimisation.

— **Loss standard (L_2^2).** Le graphique en haut à gauche montre un "bol" unique et lisse dont le minimum se situe exactement à la moyenne des données $(0, 0)$. C'est mathématiquement ce qui provoque le flou : l'optimiseur descend vers la moyenne globale, ignorant que la réalité est composée de deux groupes distincts.

— **λ = 1** : L'influence de la régularisation commence à se faire sentir. La surface s'aplatit au centre, mais la force d'attraction vers les modes réels n'est pas encore assez puissante pour briser totalement la tendance au lissage.

— **λ = 5** : Le paysage change radicalement. Deux "puits de potentiel" distincts apparaissent au-dessus des vrais modes. L'optimiseur doit maintenant "choisir" un camp. Le centre (la moyenne) devient une zone instable (une crête ou un plateau), ce qui force le modèle à générer une valeur nette appartenant à l'un des deux modes.

— **λ = 50** : Ici, l'ancrage est extrême. Les puits sont extrêmement profonds et étroits. Le modèle est "verrouillé" sur les positions des données réelles. C'est la solution proposée dans le papier VAEAC pour garantir que, même si le code latent est incertain, la reconstruction finale sera physiquement plausible (ancrée sur un mode existant) plutôt qu'une moyenne aberrante.

6.4.2 Algorithme Généralisé et Sélection de Modèle

Pour généraliser cette approche à un nombre inconnu de modes, nous proposons un algorithme en deux temps. D'abord, nous déterminons la structure de la distribution cible via un algorithme de *K-Means*. Le nombre optimal de clusters K^* est sélectionné automatiquement en maximisant le **coefficient de Silhouette**, une métrique qui évalue la cohérence des clusters (cohésion intra-classe vs séparation inter-classe).

L'algorithme complet, intégrant la découverte des modes et l'entraînement régularisé, est détaillé ci-dessous :

Algorithme 1 Entraînement VAEAC avec Régularisation Modale Adaptative

Entrée : Dataset d'entraînement $\mathcal{D}$, paramètre de régularisation λ, borne K_{max}.
Sortie : Paramètres optimisés θ^*, ψ^*, ϕ^*.
procedure INITIALISATION DES MODES
    $Scores \leftarrow []$
    for $k = 2$ **to** K_{max} **do**
        Appliquer K-Means($\mathcal{D}$, k) $\rightarrow$ Clusters $C_k = \{c_1, \ldots, c_k\}$
        Calculer $S_k \leftarrow$ SilhouetteScore($\mathcal{D}, C_k$)
        $Scores$.append(S_k)
    end for
    $K^* \leftarrow \text{argmax}_k(Scores)$ ▷ Sélection du nombre optimal de modes
    $C^* \leftarrow \{c_1, \ldots, c_{K^*}\}$ ▷ Centroïdes retenus
end procedure
procedure ENTRAÎNEMENT VAEAC
    while non convergé **do**
        Échantillonner batch (x, b)
        Calculer ELBO standard : $\mathcal{L}_{ELBO}$
        Calculer pénalité : $\Omega(x_{rec}) = \min_{c \in C^*} ||x_{rec} - c||^2$
        Loss Totale : $\mathcal{L}_{totale} = -\mathcal{L}_{ELBO} + \lambda \cdot \Omega(\mu_\theta(z))$
        Mettre à jour θ, ψ, ϕ via rétropropagation
    end while
end procedure

6.4.3 Limitations et considérations théoriques

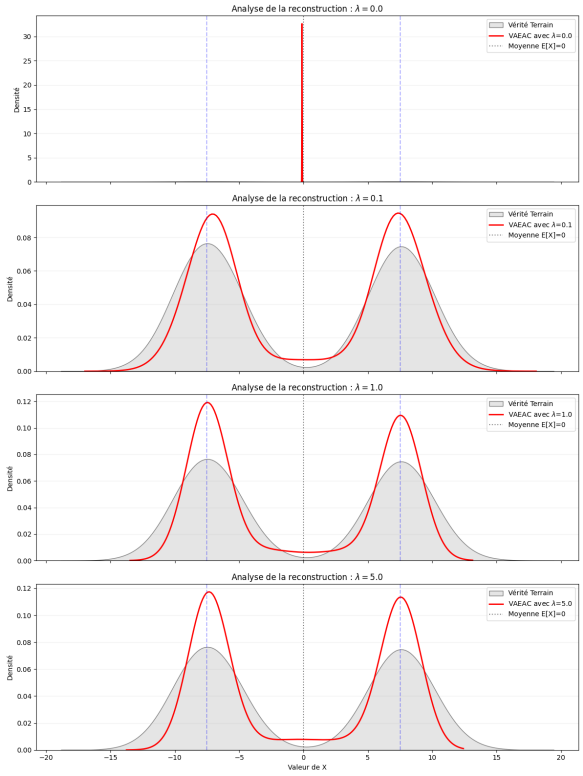
La régularisation par ancrage modal, bien qu'efficace pour produire des échantillons plausibles et éviter le lissage, introduit un biais systématique dans la modélisation de la distribution conditionnelle. En forçant les prédictions vers le mode le plus proche, elle modifie les poids du mélange dans la distribution conditionnelle $p(x_{\bar{S}} \mid x_S)$.

Considérons le cas où la vraie distribution conditionnelle est un mélange équiprobable de deux composantes : $p(x_{\bar{S}} \mid x_S) = 0.5 \cdot \mathcal{N}(-\mu, \sigma^2) + 0.5 \cdot \mathcal{N}(\mu, \sigma^2)$. La régularisation modale avec $\lambda \rightarrow \infty$ produira toujours des échantillons proches de μ ou $-\mu$ selon l'initialisation, mais perdra l'information sur l'équiprobabilité des modes.

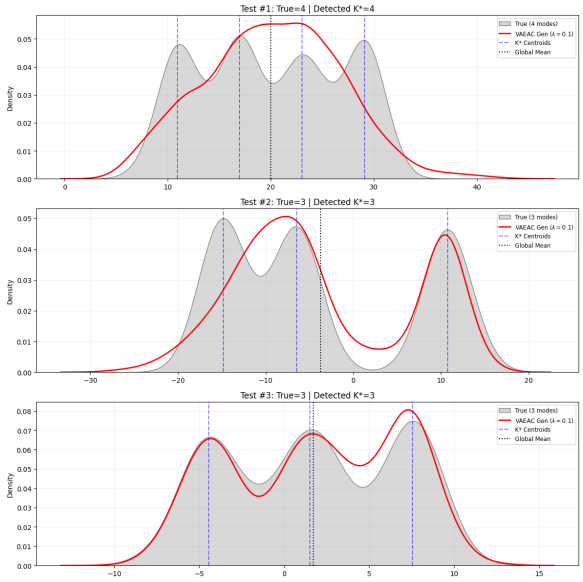
Ce biais est acceptable dans une perspective **générative** où l'objectif est de produire des échantillons réalistes, mais il doit être reconnu dans le contexte de l'**explicabilité** où l'estimation précise de l'espérance conditionnelle est primordiale. La régularisation modale représente donc un compromis entre :

— **Plausibilité des échantillons** : éviter les imputations dans les régions de faible densité
— **Fidélité distributionnelle** : préserver exactement les poids et formes des modes conditionnels

Pour les applications d'explicabilité, ce compromis peut être justifié si le modèle prédictif f est plus sensible à la position des modes qu'à leur probabilité exacte. Cependant, dans les cas où les proportions conditionnelles sont informatives (ex : classification avec probabilités de classe), une approche plus fidèle comme les mélanges de Gaussians dans le décodeur pourrait être préférable.



(a) Convergence sur une distribution bimodale symétrique (optimisation via GridSearch).



(b) Généralisation multimodale : K^* déterminé par le coefficient de Silhouette ($\lambda = 0.1$).

FIGURE 7 – **Validation de la Régularisation Modale.** La méthode parvient à ancrer les prédictions sur les modes réels plutôt que sur la moyenne (creux de densité), validant l'approche théorique.

Hyperparamètre	**Masque**	**MSE Global**	**Variance Global**
$\lambda = 0$ (Baseline)	MOITIÉ BASSE	0.058828	0.026830
$\lambda = 0.1$	MOITIÉ BASSE	0.056543	0.027868
$\lambda = 1$	MOITIÉ BASSE	**0.050437**	0.035011
$\lambda = 5$	MOITIÉ BASSE	0.053640	0.043020
$\lambda = 10$	MOITIÉ BASSE	0.053975	0.042040
$\lambda = 30$	MOITIÉ BASSE	0.054381	**0.042360**

TABLE 6 – Impact de la pondération structurelle λ sur la précision (MSE) et la netteté (Variance).

Analyse des régimes de reconstruction L'analyse des métriques globales révèle un comportement contre-intuitif mais fondamental pour la validation de la méthode. Contrairement aux hypothèses initiales suggérant qu'une forte contrainte structurelle lisserait les résultats par un effet de moyenne, la **Variance Global** augmente de manière significative avec λ (+57, 8% entre $\lambda = 0$ et $\lambda = 30$). Cela démontre que le modèle ne se contente pas de projeter l'image sur un mode flou, mais utilise le prototype pour "nettoyer" activement le fond de l'image (moyennage des zones noires), augmentant ainsi mécaniquement le contraste.

Cette dynamique permet de générer des images différenciées malgré la projection sur le mode, car le réseau sature les intensités plutôt que de les lisser.



FIGURE 8 – Évolution qualitative des reconstructions VAEAC en fonction de λ (Masque : Moitié Basse).

Les images ont été applaties, permettant ainsi de respecter l'architecture du VAEAC initial. Les changements d'architecture (CNN etc) n'ont pas réellement changer le problème.

6.5 Dérivation de l'ELBO du VAEAC

Soit $x \in \mathbb{R}^d$ un vecteur de données et $S \subset \{1, \dots, d\}$ l'ensemble des dimensions observées. On note x_S les features observées et $x_{\bar{S}}$ les features manquantes. Le VAEAC introduit une variable latente z et un réseau probabiliste conditionnel $p_\theta(x_{\bar{S}} \mid x_S, z)$ avec un encodeur $q_\phi(z \mid x_S, x_{\bar{S}})$.

1. Objectif marginal : La vraisemblance marginale des données manquantes conditionnelles aux observées s'écrit :

$$\log p_\theta(x_{\bar{S}} \mid x_S) = \log \int p_\theta(x_{\bar{S}} \mid x_S, z) p(z \mid x_S)\, dz. \tag{32}$$

Cette intégrale est intractable en pratique. On introduit l'approximation variationnelle $q_\phi(z \mid x_S, x_{\bar{S}})$.

2. Utilisation de l'inégalité de Jensen :

$$\log p_\theta(x_{\bar{S}} \mid x_S) = \log \int q_\phi(z \mid x_S, x_{\bar{S}}) \frac{p_\theta(x_{\bar{S}} \mid x_S, z) p(z \mid x_S)}{q_\phi(z \mid x_S, x_{\bar{S}})} dz \tag{33}$$

$$\geqslant \mathbb{E}_{z \sim q_\phi(z \mid x_S, x_{\bar{S}})} \Big[\log p_\theta(x_{\bar{S}} \mid x_S, z) + \log p(z \mid x_S) - \log q_\phi(z \mid x_S, x_{\bar{S}}) \Big]. \tag{34}$$

Cette inégalité de Jensen définit l'ELBO conditionnelle :

$$\mathrm{ELBO}(x_{\bar{S}} \mid x_S) = \mathbb{E}_{z \sim q_\phi(z \mid x_S, x_{\bar{S}})} \big[\log p_\theta(x_{\bar{S}} \mid x_S, z) \big] - \mathrm{KL}\big(q_\phi(z \mid x_S, x_{\bar{S}}) \| p(z \mid x_S)\big). \tag{35}$$

4. Cas particulier VAE standard : Si $S = \emptyset$, cette formulation se réduit à l'ELBO classique d'un VAE. Pour le VAEAC, l'encodeur et le décodeur sont conditionnés sur les observations x_S, ce qui permet un échantillonnage arbitraire de $x_{\bar{S}}$ en fonction de n'importe quelle observation partielle.

5. Optimisation : L'ELBO est maximisée par descente de gradient stochastique, en utilisant la reparametrization trick pour rendre l'échantillonnage différentiable :

$$z = \mu_\phi(x_S, x_{\bar{S}}) + \sigma_\phi(x_S, x_{\bar{S}}) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

Cette formulation est la base du VAEAC et garantit que l'apprentissage capture la distribution conditionnelle $p(x_{\bar{S}} \mid x_S)$.

6.5.1 Lien entre Log-Vraisemblance et MSE

Pour justifier l'utilisation de la norme L^2 dans le terme de reconstruction, nous faisons l'hypothèse que le décodeur modélise une distribution Gaussienne isotrope de **variance fixe** σ^2 (indépendante de z) centrée sur la sortie du réseau $\mu_\theta(z)$:

$$p_\theta(x|z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{||x - \mu_\theta(z)||^2}{2\sigma^2}\right) \tag{47}$$

Cette hypothèse de variance constante est courante dans les implémentations de VAE pour des raisons de stabilité numérique. Sous cette hypothèse, la maximisation de la vraisemblance devient équivalente à la minimisation de l'erreur quadratique :

$$\begin{aligned}
\theta^* &= \text{argmax}_\theta \, \mathbb{E}_z[\log p_\theta(x|z)] \\
&= \text{argmax}_\theta \, \mathbb{E}_z\left[\underbrace{-\frac{d}{2}\log(2\pi\sigma^2)}_{\text{constante}} - \frac{1}{2\sigma^2}||x - \mu_\theta(z)||^2\right] \\
&= \text{argmin}_\theta \, \mathbb{E}_z[||x - \mu_\theta(z)||^2]
\end{aligned} \tag{48}$$

Remarque importante : Dans des variantes de VAE, le décodeur peut apprendre également la variance $\sigma^2(z)$. Dans ce cas, le terme $\log \sigma^2(z)$ ne disparaît pas et la fonction de coût devient une MSE pondérée adaptative. Notre implémentation suit l'approche standard avec variance fixe, ce qui justifie l'utilisation de la perte L^2 pure (voir extension, la problématique restera la même).

7 Analyses Complémentaires et Détails Techniques

7.1 Architecture des Skip-Connections

Pour faciliter la propagation du gradient et la reconstruction fine des détails, le VAEAC intègre des connexions résiduelles. Le schéma ci-dessous illustre le flux d'information :

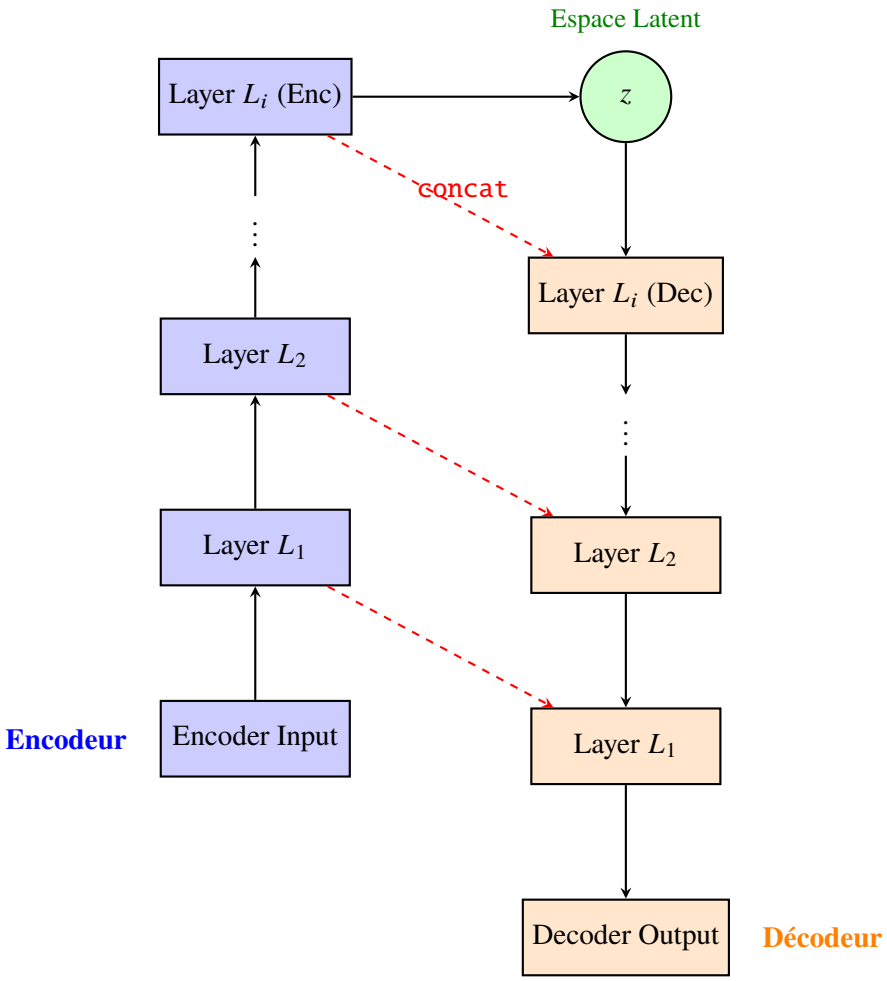


FIGURE 9 – Architecture des Skip-Connections dans le VAEAC. Les activations de chaque couche de l'encodeur sont concaténées aux entrées symétriques du décodeur, préservant l'information locale tout en permettant une compression globale via l'espace latent z.

7.2 Analyse Mathématique des Skip-Connections : Résolution du Vanishing Gradient

L'utilisation de connexions résiduelles (Skip-Connections) ne se justifie pas uniquement par la préservation de l'information locale ; elle joue un rôle structurel fondamental dans la dynamique de l'optimisation en luttant contre la disparition du gradient (*Vanishing Gradient Problem*).

Considérons un réseau profond standard. Soit x_l la sortie de la couche l. La propagation avant est donnée par $x_{l+1} = \mathcal{F}(x_l, W_l)$. Lors de la rétropropagation, le gradient de la fonction de perte $\mathcal{L}$ par rapport à une couche initiale x_l s'obtient par la règle de la chaîne (Chain Rule) :

$$\frac{\partial \mathcal{L}}{\partial x_l} = \frac{\partial \mathcal{L}}{\partial x_L} \cdot \prod_{i=l}^{L-1} \frac{\partial x_{i+1}}{\partial x_i} = \frac{\partial \mathcal{L}}{\partial x_L} \cdot \prod_{i=l}^{L-1} W_i \cdot \sigma'(z_i) \tag{36}$$

Dans cette configuration multiplicative, si les valeurs singulières des matrices de poids W_i sont faibles (ou si la dérivée de l'activation $\sigma' < 1$), le produit tend exponentiellement vers zéro à mesure que la profondeur $(L - l)$ augmente. Cela empêche la mise à jour des poids dans les premières couches.

L'apport du bloc Résiduel Avec une Skip-Connection, la relation devient $x_{l+1} = x_l + \mathcal{F}(x_l, W_l)$. Le gradient local devient alors :

$$\frac{\partial x_{l+1}}{\partial x_l} = I + \frac{\partial \mathcal{F}}{\partial x_l} \tag{37}$$

où I est la matrice identité. En réinjectant ce terme dans l'équation de rétropropagation, le gradient transmis à la couche l devient :

$$\frac{\partial \mathcal{L}}{\partial x_l} = \frac{\partial \mathcal{L}}{\partial x_{l+1}} \cdot \left(I + \frac{\partial \mathcal{F}}{\partial x_l} \right) = \underbrace{\frac{\partial \mathcal{L}}{\partial x_{l+1}}}_{\text{Terme Identité}} + \underbrace{\frac{\partial \mathcal{L}}{\partial x_{l+1}} \cdot \frac{\partial \mathcal{F}}{\partial x_l}}_{\text{Terme Résiduel}} \tag{38}$$

Cela permet d'entraîner des réseaux très profonds (comme notre VAEAC) sans que les couches d'encodage initiales ne souffrent d'une absence de mise à jour.

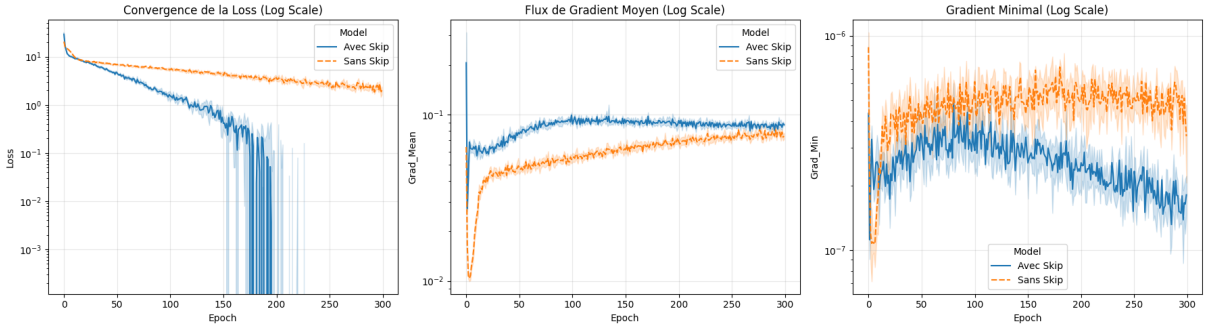


FIGURE 10 – **Impact des Skip-Connections sur l'optimisation (échelle log).** À gauche : convergence de la loss ; au centre : flux de gradient moyen ; à droite : gradient minimal, comparant les modèles avec et sans skip-connections.

Analyse. La figure 10 met en évidence trois effets cohérents des skip-connections :
— (i) une **convergence plus rapide** de la loss (courbe bleue) que sans skip, indiquant une optimisation plus efficace ;
— (ii) un **gradient moyen** plus élevé et plus stable, ce qui traduit une meilleure propagation du signal d'erreur vers les premières couches ;
— (iii) un **gradient minimal** globalement moins sujet à l'effondrement, symptôme direct d'une atténuation du *vanishing gradient*.
pratique, ces observations justifient l'usage de skip-connections dans le VAEAC pour stabiliser l'entraînement et améliorer la qualité des reconstructions.

Remarque sur le rôle des SkipBlocks. L'introduction des connexions résiduelles au sein des blocs d'architecture (SkipBlocks) modifie fondamentalement le flux d'information. Sans ces connexions, le modèle serait contraint de compresser l'intégralité de l'information des données observées x_S dans le vecteur latent z. Cette compression risque d'effacer les détails précis nécessaires à une reconstruction fidèle.Les SkipBlocks permettent de résoudre ce problème en créant un lien direct entre les couches de l'encodeur et celles du décodeur. On observe alors une répartition naturelle des tâches :
— Le vecteur latent z se charge de capturer la structure globale et les dépendances complexes (l'allure générale).
— Les connexions résiduelles transmettent directement les caractéristiques brutes de l'entrée x_S (les détails).
Ainsi, le modèle n'a pas besoin de "mémoriser" les détails exacts de l'entrée dans le code latent, ce qui facilite l'apprentissage et améliore la cohérence de l'imputation.

7.3 Analyse de la sensibilité aux choix des hyperparamètres

Paramètre	Valeur	Description
Optimiseur	AdamW	Avec découplage du weight decay
Learning Rate	$1e^{-3}$	Décroissance exponentielle ($\gamma = 0.95$)
Batch Size	64	Compromis vitesse/stabilité
Profondeur DLN	4	Nombre de matrices diagonales (Réseau Student)
Patience	15	Critère d'arrêt (Early Stopping)
Dimension Latente	8	Taille de l'espace compressé z

TABLE 7 – Récapitulatif des hyperparamètres retenus après validation croisée.

Grille explorée (dimension latente et pas d'apprentissage). Les Tables 8 et 9 résument une recherche sur grille sur plusieurs jeux de données, en faisant varier la dimension latente $L \in \{4, 8, 16, 32\}$ et le *learning rate*.

Dataset	$L = 4$	$L = 8$	$L = 16$	$L = 32$
abalone	0.0001	0.0001	0.0050	0.0001
california	0.0001	0.0001	0.0010	0.0001
copula	0.0001	0.0001	0.0010	0.0001
diabetes	0.0050	0.0010	0.0050	0.0010
gmm	0.0050	0.0050	0.0010	0.0050
iid	0.0050	0.0050	0.0050	0.0050

TABLE 8 – Valeurs du *learning rate* retenues (par dataset) pour chaque dimension latente.

Dataset	$L = 4$	$L = 8$	$L = 16$	$L = 32$
abalone	0.937678	0.958579	0.928562	0.932899
california	0.630390	0.631549	0.621342	0.611744
copula	0.610468	0.590914	0.641578	0.637480
diabetes	0.350620	0.320952	0.334418	0.331744
gmm	0.906764	0.909935	0.881104	0.916204
iid	-0.009247	-0.006321	-0.009822	-0.001979

TABLE 9 – Performance R^2 moyenne sur 5 essai en fonction de la dimension latente (meilleure valeur en gras par ligne).

Analyse. Globalement, $L = 8$ apparaît comme un bon compromis sur des données tabulaires réelles (*abalone* : $R^2 = 0.9586$, *california* : $R^2 = 0.6315$), tandis que $L = 16$ est préférable lorsque la structure de dépendance est plus marquée (*copula* : $R^2 = 0.6416$). Pour un mélange gaussien, l'augmentation de capacité aide (*gmm* : meilleur à $L = 32$ avec $R^2 = 0.9162$), ce qui est cohérent avec une multimodalité plus difficile à capturer avec un latent trop petit.

Concernant l'optimisation, on observe que des *learning rates* faibles (10^{-4}) suffisent sur *abalone/california/copula* (stabilité de l'entraînement), alors que *diabetes/gmm/iid* nécessitent des pas plus élevés (jusqu'à 5×10^{-3}) pour éviter une convergence trop lente.

Enfin, le cas *iid* illustre une limite : les R^2 proches de 0 (voire négatifs) indiquent qu'augmenter L ne compense pas une structure latente peu informative dans ce protocole (ou un modèle sous-spécifié pour ce dataset), d'où l'intérêt de sélectionner L et le pas d'apprentissage de manière dépendante du contexte.

Remarque 6. Le fait que le VAEAC obtienne un score proche de 0 (la moyenne) montre qu'il n'hallucine pas de relations inexistantes. C'est une preuve de robustesse contre l'overfitting, les données observées n'apporte aucune information, la moyenne des données sera renvoyés $v(\emptyset) = \mathbb{E}(f(x))$

Bibliographie

— **Chalkiadakis, G., Elkind, E., & Wooldridge, M. (2011).** *Computational Aspects of Cooperative Game Theory*. Morgan & Claypool Publishers.

— **Ivanov, O., Figurnov, M., & Vetrov, D. (2019).** Variational autoencoder with arbitrary conditioning. *arXiv preprint arXiv :1806.02382v3*. https://arxiv.org/abs/1806.02382

— **Le Corff, S. (2025).** *Méthodes de simulation* (support de cours). Master 2 Statistique, Sorbonne Université.

— **Lundberg, S. M., & Lee, S.-I. (2017).** A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

— **MonshotData.** Le danger de l'imputation par la moyenne ou par zéro. Blog. https://www.monshotdata.com/p/imputation-knn-valeurs-manquantes

— **Olsen, L. H. B., Glad, I. K., Jullum, M., & Aas, K. (2022).** Using Shapley values and variational autoencoders to explain predictive models with dependent mixed features. Source principale. https://arxiv.org/abs/2111.13507

— **Scikit-Learn Developers.** Selecting the number of clusters with silhouette analysis on KMeans clustering. Documentation officielle. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

— **Stekhoven, D. J., & Bühlmann, P. (2012).** MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* https://arxiv.org/abs/1105.0828.

— **Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, Ole Winther(2016).** Autoencoding beyond pixels using a learned similarity metric. *Bioinformatics*.

— **Pourquoi les images créées par les GAN sont-elles plus nettes que les images créées par les VAE ?, Reddit** en cliquant ici

Extension : Traces de recherche

A Cas général : Décodeur avec variance apprise

A.1 Formulation du problème

Considérons un décodeur gaussien probabiliste $p_\theta(x_{\bar{S}} \mid z, x_S)$ qui prédit à la fois la moyenne conditionnelle $\mu_\theta(z)$ et la variance $\sigma^2_\theta(z)$. Pour un code latent z et une observation partielle x_S donnés, l'objectif d'optimisation (Log-Vraisemblance Négative) s'écrit :

$$\mathcal{L}(\theta) = \mathbb{E}_{x_{\bar{S}} \sim p(x_{\bar{S}} | x_S)} \left[\log \sigma^2_\theta(z) + \frac{(x_{\bar{S}} - \mu_\theta(z))^2}{\sigma^2_\theta(z)} \right]$$

A.2 Solution optimale théorique

En minimisant cette fonction de perte par rapport aux sorties du réseau, les conditions du premier ordre donnent :

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu_\theta} &= \mathbb{E}\left[\frac{-2(x_{\bar{S}} - \mu_\theta)}{\sigma^2_\theta} \right] = 0 \\
&\Rightarrow \mu^*_\theta(z) = \mathbb{E}[x_{\bar{S}} \mid x_S] \\
\frac{\partial \mathcal{L}}{\partial \sigma^2_\theta} &= \mathbb{E}\left[\frac{1}{\sigma^2_\theta} - \frac{(x_{\bar{S}} - \mu_\theta)^2}{\sigma^4_\theta} \right] = 0 \\
&\Rightarrow \sigma^{*2}_\theta(z) = \mathbb{E}\left[(x_{\bar{S}} - \mu^*_\theta(z))^2 \mid x_S \right] = \mathrm{Var}(x_{\bar{S}} \mid x_S)
\end{aligned}$$

A.3 Application au cas bimodal symétrique

Reprenons le cas critique d'une distribution conditionnelle bimodale $p(x_{\bar{S}} \mid x_S)$ composée de deux gaussiennes de poids égaux centrées en $-\mu_0$ et μ_0 avec une variance intrinsèque σ^2_0. Le décodeur optimal convergera vers :

$$\begin{aligned}
\mu^*_\theta(z) &= 0 \quad \text{(Moyenne globale, lissage)} \\
\sigma^{*2}_\theta(z) &= \mathbb{E}[x^2_{\bar{S}} \mid x_S] - (\mu^*_\theta)^2 \\
&= (\mu^2_0 + \sigma^2_0) - 0 \\
&= \mu^2_0 + \sigma^2_0
\end{aligned}$$

A.4 Interprétation

Cette analyse démontre que l'apprentissage de la variance (σ^2_θ) ne résout pas le problème structurel du lissage induit par l'unimodalité de la gaussienne de sortie :

1. La **prédiction moyenne** μ^*_θ reste "bloquée" sur l'espérance conditionnelle (0), au milieu des modes (zone de faible densité).
2. La **variance apprise** σ^{*2}_θ gonfle pour englober l'incertitude épistémique (la distance entre les modes μ^2_0) en plus de l'incertitude aléatorique (σ^2_0).
3. Le modèle produit une *unique* gaussienne très large, couvrant les deux modes sans pouvoir les séparer, ce qui confirme la pertinence de l'analyse simplifiée à variance fixe présentée précédemment.

B Du VAEAC au DAC : du VAE à la diffusion conditionnelle

Cette section propose une lecture unifiée des modèles génératifs utilisés pour l'imputation conditionnelle. L'évolution conceptuelle consiste à passer (i) d'une compression en une étape unique via un espace latent (VAE/VAEAC) à (ii) une modélisation par une cascade de transitions latentes (diffusion), permettant (iii) un apprentissage conditionnel fin où le modèle tente de recouvrer la trajectoire du processus direct sachant les données observées.

Notations. On note $x_0 \in \mathbb{R}^d$ une observation complète (donnée réelle), x_t la variable bruitée à l'étape $t \in \{1, \dots, T\}$, et I la matrice identité. Le bruit standard est $\epsilon \sim \mathcal{N}(0, I)$. Les coefficients de diffusion vérifient $\alpha_t \in (0, 1)$, $\beta_t = 1 - \alpha_t$ et $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

B.1 A.1 Point de départ : VAE et log-vraisemblance (ELBO)

Dans un VAE standard, on cherche à maximiser la vraisemblance marginale $\log p(x)$. Comme elle est intractable, on maximise une borne inférieure variationnelle (ELBO) :

$$\log p(x) \geq \mathbb{E}_{q(z|x)}[\log p(x \mid z)] - D_{KL}\big(q(z \mid x) \,\|\, p(z)\big). \tag{39}$$

Cette formulation repose sur un saut unique vers une variable latente z : l'information est encodée (compressée) puis décodée (reconstruite) en une seule étape stochastique.

B.2 A.2 Évolution : modèles de diffusion comme "cascade de latents"

Les modèles de diffusion généralisent cette approche en introduisant une séquence de variables latentes $x_1, \dots, x_T$ de même dimension que x_0. Le processus direct (bruitage) est fixé et markovien :

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}\, x_{t-1},\, \beta_t I\big), \qquad t = 1, \dots, T. \tag{40}$$

Grâce aux propriétés gaussiennes, la distribution marginale à l'étape t s'écrit directement :

$$q(x_t \mid x_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, x_0,\, (1 - \bar{\alpha}_t) I\big). \tag{41}$$

La log-vraisemblance $\log p_\theta(x_0)$ est minorée par une ELBO qui se décompose non plus en une reconstruction unique, mais en une somme de termes d'alignement le long de la trajectoire :

$$\log p_\theta(x_0) \geq \mathbb{E}_{q(x_1|x_0)}\big[\log p_\theta(x_0 \mid x_1)\big] - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}\Big[D_{KL}\big(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\big)\Big]. \tag{42}$$

Contrairement au VAE, la diffusion apprend une suite de transformations locales inversant progressivement le bruit.

B.3 A.3 Conditionnement arbitraire et imputation

Dans le cadre de l'imputation, l'objectif est de modéliser $p(x_{\bar{S}} \mid x_S)$, où S désigne les indices observés et $\bar{S}$ les indices manquants. Avec un masque binaire $m \in \{0, 1\}^d$, on a :

$$x_S = x \odot m, \qquad x_{\bar{S}} = x \odot (1 - m). \tag{43}$$

Une approche naïve consisterait à utiliser un modèle inconditionnel et à injecter x_S *a posteriori* lors de l'échantillonnage. Cependant, pour garantir la cohérence des jointures entre parties observées et générées, le modèle doit apprendre explicitement ce conditionnement.

B.4 A.4 Diffusion avec conditionnement arbitraire (DAC) : Alignement Variationnel

Pour réaliser une imputation robuste, le processus inverse est paramétré pour dépendre explicitement des données observées x_S :

$$p_\theta(x_{t-1} \mid x_t, x_S) = \mathcal{N}\big(\mu_\theta(x_t, t, x_S),\, \Sigma_\theta(x_t, t, x_S)\big). \tag{44}$$

L'objectif d'entraînement s'interprète alors comme un **alignement de distributions conditionnelles**. À chaque étape t, le modèle p_θ (qui ne voit que x_S et x_t) doit approcher au mieux la distribution *a posteriori* réelle du processus direct q (qui, lui, a accès à l'information complète x_0 par construction).

On minimise la divergence suivante sur les composantes manquantes :

$$L_{\text{cond}} = \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \Big[D_{KL}\big(\underbrace{q(x_{t-1,\bar{S}} \mid x_t, x_0)}_{\text{Cible : issue de } x_0 \text{ complet}} \;\Big\|\; \underbrace{p_\theta(x_{t-1,\bar{S}} \mid x_t, x_S)}_{\text{Modèle : conditionné par } x_S} \big) \Big]. \tag{45}$$

Remarque 7 (Guidage par la trajectoire). Cette formulation met en évidence que l'apprentissage consiste à "guider" le modèle vers la trajectoire de débruitage idéale (celle qu'on aurait suivie si l'on connaissait x_0). Le modèle apprend à halluciner de manière cohérente le chemin de $x_{t,\bar{S}}$ vers $x_{t-1,\bar{S}}$ en s'appuyant sur les indices fournis par x_S.

Remarque 8 (Lien avec le VAEAC). Le cas $T = 1$ constitue un cas limite dégénéré. Si le processus se réduit à une seule étape, le modèle apprend à prédire directement $x_{\bar{S}}$ à partir d'une version bruitée et de x_S. Bien que cela rappelle structurellement le VAEAC (encodage/décodage conditionnel), la différence majeure réside dans l'absence d'apprentissage d'un espace latent compressé : ici, l'espace latent a la même dimension que les données et le prior est fixé.

B.5 Analyse des Résultats

Cette section présente une comparaison quantitative et qualitative entre l'approche *Variational Autoencoder with Arbitrary Conditioning* (VAEAC) et la méthode de diffusion *Diffusion with Abstract Conditioning* (DAC) pour $T = 5$ et $T = 10$ pas de temps. L'évaluation est réalisée sur trois types de masques : Vertical (occultation inférieure), Latéral (occultation droite) et Damier (perte de 50% des pixels).

B.6 Analyse Quantitative et Comparative

Le Tableau ci-dessus présente les métriques de reconstruction (RMSE, MAE, PSNR) pour les trois configurations (VAEAC, DAC $T = 5$, DAC $T = 10$) soumises aux trois typologies de masques. L'analyse des données met en lumière trois dynamiques fondamentales :

B.6.1 La Suprématie de la Diffusion sur l'Information Dispersée (Masque Damier)

B.6.2 Métrique de Qualité : Peak Signal-to-Noise Ratio (PSNR)

Pour évaluer la fidélité des reconstructions, nous utilisons le *Peak Signal-to-Noise Ratio* (PSNR), exprimé en décibels (dB). Il est défini par la formule suivante :

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{46}$$

où MAX_I est la valeur maximale possible d'un pixel (ici 1.0) et MSE est l'erreur quadratique moyenne sur les pixels masqués.

Pourquoi cette métrique ? L'utilisation du PSNR est privilégiée ici pour deux raisons :

TABLE 10 – Comparaison quantitative des performances de reconstruction sur le jeu de test MNIST (Moyenne ± Écart-type). Les meilleures performances par catégorie sont indiquées en gras.

Masque	Modèle	RMSE (↓)	MAE (↓)	PSNR (dB) (↑)
Damier	VAEAC	0.257 ± 0.035	0.150 ± 0.020	11.86 ± 1.13
	DAC ($T = 5$)	**0.143 ± 0.032**	**0.061 ± 0.019**	**17.15 ± 2.19**
	DAC ($T = 10$)	0.155 ± 0.034	0.069 ± 0.021	16.45 ± 2.15
Latéral	VAEAC	0.265 ± 0.035	0.159 ± 0.022	11.59 ± 1.12
	DAC ($T = 5$)	**0.218 ± 0.059**	**0.103 ± 0.038**	**13.60 ± 2.65**
	DAC ($T = 10$)	0.220 ± 0.057	0.107 ± 0.038	13.47 ± 2.48
Vertical	VAEAC	0.265 ± 0.039	0.159 ± 0.024	11.64 ± 1.24
	DAC ($T = 5$)	0.231 ± 0.062	**0.108 ± 0.040**	13.08 ± 2.53
	DAC ($T = 10$)	**0.228 ± 0.060**	0.110 ± 0.039	**13.19 ± 2.52**

1. **Échelle Logarithmique :** le PSNR utilise une échelle logarithmique. Cela permet de mieux distinguer les améliorations de qualité fines (comme la netteté des bords) qui sont cruciales dans notre comparaison entre le flou du VAEAC et la netteté du DAC.
2. **Standard Industriel :** C'est la métrique de référence en restauration d'image, permettant de comparer objectivement le rapport entre le signal utile (la structure du chiffre) et le bruit (les erreurs de reconstruction).

L'écart de performance est le plus flagrant sur le masque de type « Damier ».

— **Performance du DAC :** Le modèle *DAC ($T = 5$)* atteint un PSNR record de **17.15 dB** et une erreur RMSE minimale de **0.143**. Cela démontre l'excellente capacité des modèles de diffusion à exploiter les corrélations locales (voisinage de pixels) pour propager l'information, agissant comme un algorithme de *super-resolution* très efficace.

— **Limitation du VAEAC :** En comparaison, le VAEAC plafonne à **11.86 dB** de PSNR. L'architecture goulot d'étranglement du VAE tend à filtrer les hautes fréquences nécessaires pour reconstruire la grille fine du damier, résultant en une perte massive de détails (RMSE de 0.257).

B.6.3 L'Efficacité Inattendue de la Diffusion Rapide ($T = 5$ vs $T = 10$)

Contrairement à l'intuition classique selon laquelle plus d'étapes de diffusion améliorent la qualité, les résultats montrent une **saturation, voire une légère régression** à $T = 10$ pour ce dataset spécifique :

— Sur le masque Damier, le DAC $T = 5$ surpasse significativement le $T = 10$ (RMSE 0.143 contre 0.155).

— Sur les masques structurels (Latéral/Vertical), les performances sont statistiquement équivalentes (ex : RMSE Latéral ≈ 0.22).

Interprétation : Pour des images à faible complexité comme MNIST (binaire, géométrie simple), le *manifold* des données est atteint rapidement. Un nombre d'étapes trop élevé ($T = 10$) sur une tâche d'imputation contrainte (où une partie de l'image est fixée) peut introduire une dérive ou accumuler des erreurs d'approximation sans apporter de raffinement sémantique supplémentaire. Le modèle $T = 5$ représente ici un optimum coût/qualité.

B.6.4 Compromis Stabilité (VAEAC) vs Réalisme (DAC)

L'analyse des écarts-types ($\pm\sigma$) confirme la différence de nature entre les deux approches :

— **VAEAC (Faible Variance) :** Avec un écart-type constant et faible ($\sigma_{RMSE} \approx 0.035$ à 0.040), le VAEAC est un modèle "prudent". Il converge systématiquement vers la moyenne statistique, minimisant le risque de grosse erreur au prix d'un résultat flou.
— **DAC (Forte Variance) :** Les modèles de diffusion présentent une variance presque double ($\sigma_{RMSE} \approx 0.060$ pour le masque vertical). Cela reflète leur nature générative : face à une ambiguïté (ex : bas d'un '8' masqué), le DAC "choisit" une forme distincte. Si ce choix correspond à la vérité terrain, l'erreur est nulle ; sinon, elle est élevée. Le DAC privilégie la forme à la fidélité pixel-par-pixel moyenne.

B.6.5 Conclusion Partielle

Sur MNIST, l'approche DAC avec une diffusion courte ($T = 5$) s'impose comme la solution la plus performante, offrant un gain de **+5.5 dB** en PSNR sur les tâches d'inpainting complexe (Damier) par rapport au VAEAC. Cependant, la plus grande instabilité du DAC (écart-type élevé) sur les masques structurels (Vertical/Latéral) rappelle que l'ancrage modal reste un défi pour les modèles génératifs purement probabilistes lorsqu'une part critique de l'information structurelle est absente.

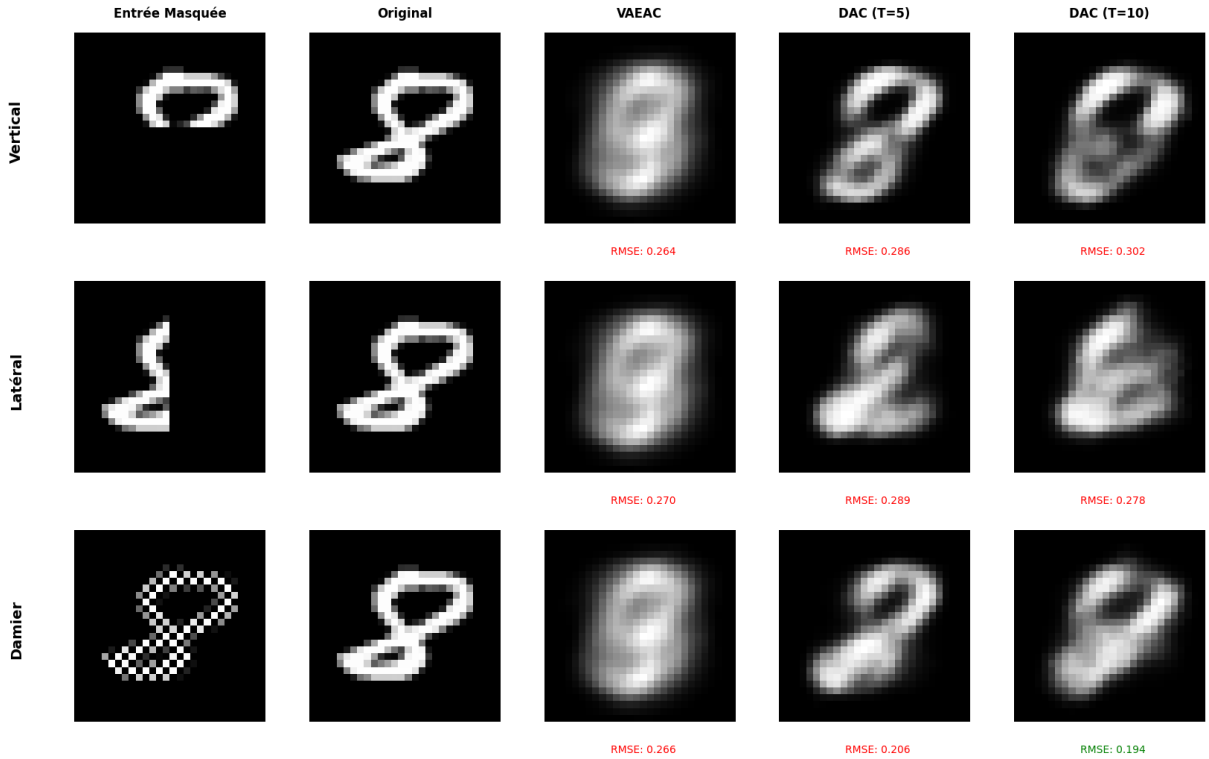


FIGURE 11 – **Comparaison visuelle et métrique de l'imputation (VAEAC vs DAC).** Reconstruction du chiffre "8" sous trois scénarios de masquage (Vertical, Latéral, Damier). Les résultats comparent le VAEAC classique et le modèle DAC avec $T = 5$ et $T = 10$ pas de diffusion. *Détails expérimentaux :* Entraînement sur **10 époques** avec une taille de lot (*batch size*) de **64**. Ces résultats sont issus d'une **architecture générale légèrement modifiée** et utilisent une **fonction de perte adaptée**. Les scores RMSE (en rouge/vert) indiquent la qualité de reconstruction pixel à pixel.

C Travaux connexes, limitations et perspectives

Ce travail a permis de valider l'approche VAEAC. Plusieurs axes de recherche restent néanmoins ouverts pour étendre ces résultats.
— **Passage à l'échelle et Parallélisation :** Les expérimentations actuelles ont été limitées par les ressources de calcul (Google Colab). L'utilisation d'architectures distribuées (multi-GPU) permettrait

d'explorer des réseaux plus profonds et de traiter des jeux de données de haute dimension, inaccessibles aux méthodes de Monte Carlo actuelles sans optimisation CUDA spécifique.

— **Modèles Génératifs de Nouvelle Génération :** Bien que le VAE reste une référence, l'état de l'art a évolué vers les *Score-Based Diffusion Models* (SBDM) et les *Flow Matching Models*. Implémenter un conditionnement arbitraire sur ces architectures pourrait offrir une meilleure couverture des modes et une fidélité accrue par rapport aux VAEs, réduisant potentiellement le besoin de régularisation modale.
— **Benchmarking Étendu :** Il serait pertinent de comparer la méthode DLN proposée (Réseau Student) avec d'autres estimateurs de l'état de l'art tels que *KernelSHAP* ou *TreeSHAP*, mais également d'explorer d'autres concepts de la théorie des jeux coopératifs.

Difficultés rencontrées

— Complexité calculatoire, certains de mes tests pouvaient prendre aux alentours d'une heure.
— Recoder l'architecture proposé, et en reproduire une qui offre des résultats assez satisfaisant.
— comprendre la complexité de l'architecture.
— comprendre les impacts d'une mauvaise simulation sur l'XAI.

C.0.1 Quelques test disponibles ici pour le reproductibilité des résultats et la vérification de l'architecture

— quelques tests disponibles ici
— Du VAEAC au DAC, un test sur MNIST ici