SUCCESSIVE-REFINEMENT LATENT VARIABLE GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many generative models construct samples progressively, through sequences of latent variables, denoising steps, or iterative refinements. We study when this process forms a meaningful hierarchy, in which each stage reveals new information, improves the current reconstruction, and enables coarse-to-fine control. Our focus is on hierarchical latent variable models (HLVMs), where multiple latent variables concur to build a coarse-to-fine generation pipeline. We analyze them through *successive refinement*, the classical theory of layered lossy coding, which asks whether every partial representation achieves the smallest reconstruction error permitted by the information it carries about the source. This view yields two core principles: I) control the cumulative information rate available at each stage; II) supervise every intermediate reconstruction. We design a novel objective for hierarchical models following these principles, and show it recovers the best model-realizable hierarchy at the achieved information rates. We then analyze two canonical HLVMs under the same criteria. We find the ELBO of hierarchical variational autoencoders (HVAEs) does not incentivize hierarchical representations, as it constrains only the total rate and final reconstruction, whereas denoising diffusion probabilistic models (DDPMs) realize a special case in which the encoder is fixed, and rate–distortion optimal only for Gaussian sources under MSE. We instantiate the proposed objective in an HVAE baseline, the Successive Refinement HVAE (SR-HVAE). On analytically tractable sources it matches every prediction of the SR theory. On images it makes every latent active, enables reaching rate regions the standard ELBO never targets, and improves overall generative ability.

1 INTRODUCTION

Modern image generation increasingly relies on hierarchical and iterative construction. Hierarchical variational autoencoders (HVAEs) organize representations across stochastic latent levels (Sønderby et al., 2016; Vahdat & Kautz, 2020; Child, 2020); diffusion models (DDPM) construct samples through iterative denoising (Sohl-Dickstein et al., 2015; Ho et al., 2020); StyleGANs inject latent information at multiple growing resolutions (Karras et al., 2019; Sauer et al., 2022); recent autoregressive models generate progressively finer token maps (Tian et al., 2024). Ideally, when building such hierarchical models, each stage should actively participate in the process, preserving what has been previously created, adding new information, and passing a better output forward. The common intuition is simple yet appealing: establish global structure first, then progressively refine it with local detail. Realizing it would decompose reconstruction into progressively finer problems and make it possible to model the generated output at different levels of detail. Yet a multistage architecture alone does not guarantee this behavior. A latent level may be ignored entirely, as in posterior collapse (Chen et al., 2017; Sønderby et al., 2016); or unevenly contribute to the final sample quality (Nichol & Dhariwal, 2021; Choi et al., 2022), yielding a plausible but under-expressed final sample with poor intermediate representations. We therefore ask: *how can we train a generative model so that every stage actively adds new information, improving reconstructions?*

We recast this question as a progressive communication problem, where a message is transmitted in multiple stages (Cover & Thomas, 2006; Equitz & Cover, 1991). Under this framework, the transmitted message represents the latent variable of generative models, and the amount of information encoded within it is denoted rate (R). As in generative architectures, the decoder interprets the message to deliver a reconstruction, whose distortion (D) can be quantified with respect to the source,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

i.e. the encoder's input. Rate-distortion theory defines the minimum distortion ($D(R)$) attainable at each rate, forming a frontier on which information is used optimally (Shannon, 1959; Cover & Thomas, 2006). *Successive Refinement* (SR) (Equitz & Cover, 1991; Rimoldi, 1994) studies the conditions under which a multi-step message can reach such frontier after every stage, forming a natural analogy with the hierarchical latent variable generative models (HLVMs) we study in this paper. Notably, SR proves that optimal hierarchical reconstructions are possible only for certain source–distortion pairs: some settings admit such hierarchies, whereas others must depart from the frontier at one or more stages (Chow & Berger, 1997).

Guided by SR theory, we develop a general framework for optimal information allocation in multistage generative modeling. It applies to any architecture defining a sequence of representations, a cumulative information rate, and a *prefix* reconstruction at every stage (*i.e*. the output obtainable from the first k latent variables alone). The framework yields two core principles: (i) control the cumulative rate at each stage; and (ii) supervise the reconstruction produced by every prefix, rather than only the final output. We propose a novel objective which enables training HLVMs according to these principles, so that every latent level improves the reconstruction available at its stage. Analyzing two reference HLVM families under this paradigm, we find that HVAE's ELBO does not incentivize hierarchical representations, since it constrains only the total variational rate and the final reconstruction. DDPMs instead supervise every prefix, and can be read as the special case of our objective in which the encoder is fixed, and rate–distortion optimal only for Gaussian sources under MSE.

The proposed objective is instantiated in a top-down HVAE that we train to pursue information-optimal intermediate representations, dubbed *Successive-Refinement Hierarchical Variational Autoencoder* (SR-HVAE). SR-HVAE is not a new architecture but a minimal vehicle for the objective: it retains the standard HVAE parameterization and changes only the optimization. We test it on two analytical sources with known rate–distortion structure and on two image datasets, MNIST and CelebA. The analytical sources act as a theoretical probe: the objective recovers the hierarchy SR prescribes when the source admits one, and leaves the frontier exactly where the theory says every model must. On images, the SR analysis of existing HLVMs explains *why they should pursue a meaningful hierarchy*. HVAE's ELBO is a perfect example of an objective enforcing neither of our principles. It cannot express which rate region to target, nor how to distribute the latents within it, so HVAEs settle on an allocation of their own that uses only a fraction of what a prescribed schedule can ask for. DDPMs instead implicitly target a much wider region through their noise schedule, an advantage our objective makes explicit. ELBO also leaves intermediate reconstructions unsupervised, making most prefixes unused: on CelebA, 12 of 32 prefixes of HVAE reconstruct no better than the data mean, a failure our objective removes entirely. Hence, the proposed objective makes each prefix active, explicit targets a prescribed rate region and decides how much of the budget is spent on coarse *vs*. fine structure. Raising the total rate budget alone moves SR-HVAE's FID from 22.8 to 13.6 on MNIST, while choosing how that budget is spread buys a further 18%, down to 11.1.

We further show that SR-HVAE and DDPMs both purse the same underlying objective, and what separates them is the encoder that realizes it. A learned encoder adapts to the source, but is trained against a variational rate: the true rate plus the divergence between the generative prior and the distribution the encoder induces (Hoffman & Johnson, 2016; Rosca et al., 2018; Tomczak & Welling, 2018; Dai & Wipf, 2019). A prescribed encoder carries no such divergence, but is rate–distortion optimal only for the source it was designed around. Experimental evidence suggests the best choice is a learned encoder with a tight bound: when decoded from its own posterior, SR-HVAE surpasses an equal-depth DDPM on both datasets, outlining the superiority of the learned hierarchy when the variational and true rates coincide. Refitting the prior alone recovers 60% of that distance on MNIST and 28% on CelebA, confirming the remaining gap sits in the prior rather than in the objective.

Our contributions are threefold. I) We re-cast hierarchical models as a progressive communication problem, and use successive refinement theory to derive a principled objective for truly hierarchical representations. II) Within this view, we show that DDPMs realize the special case of our objective in which the encoder is fixed, and rate–distortion optimal only for Gaussian sources under MSE, whereas the HVAE ELBO does not incentivize hierarchical representations at all. III) We instantiate the objective in SR-HVAE and test it on analytical and image sources. On the former, our objective matches every theoretical prediction. On the latter, we show that pursuing a meaningful hierarchy translates directly into better generative metrics, and that what then separates two instances of the same objective is the prior mismatch alone, which can be reduced without touching the hierarchy.

2 BACKGROUND

This sections provides the key concepts necessary to follow the discussion in Section 3. It fixes the notation and recalls main results of previous works, including Rate–Distortion and Successive Refinement (Shannon, 1959; Equitz & Cover, 1991), the definition of hierarchical latent variable models (HLVM) (Shu & Ermon, 2022), and basic notions of HVAEs and DDPMs. An extended background section is included in Appendix B, related works are discussed in Appendix A.

Progressive latent notation. Let $\mathbf{X} \in \mathbb{R}^m$ be a sample from a source distribution and let T denote the number of stages in a hierarchical model. We index latent variables $\mathbf{Z}_1, \dots, \mathbf{Z}_T$ in *coarse-to-fine* order, from $k = 1$ (coarsest) to $k = T$ (finest). This ordering is opposite to the usual DDPM convention, but it aligns diffusion states with HVAEs and successive refinement. The sequence $\mathbf{Z}_{1:k} := (\mathbf{Z}_1, \dots, \mathbf{Z}_k)$ indicates the latents up to stage k; $\hat{\mathbf{X}}^{(k)}$ is the corresponding reconstruction.

Rate–distortion and successive refinement. Mutual information $I(\mathbf{X}; \hat{\mathbf{X}})$ measures how much information about the source $\mathbf{X}$ is retained in an arbitrary reconstruction $\hat{\mathbf{X}}$. Given a distortion measure $d(\mathbf{X}; \hat{\mathbf{X}})$, rate–distortion theory characterizes the smallest rate R required for distortion D, or equivalently the smallest distortion D allowed by rate R (Shannon, 1959; Cover & Thomas, 2006):

$$R(D) := \inf_{\mathbb{E}[d] \le D} \frac{I(\mathbf{X}; \hat{\mathbf{X}})}{m \log 2} \text{ [bits/dim]}, \quad D(R) := \inf_{\frac{I(\mathbf{X};\hat{\mathbf{X}})}{m \log 2} \le R} \mathbb{E}[d] \text{ [MSE/dim]}, \quad d := \tfrac{\|\mathbf{X} - \hat{\mathbf{X}}\|_2^2}{m}. \tag{1}$$

A conditional distribution $p(\hat{\mathbf{X}}|\mathbf{X})$ attaining either infimum is an *RD-optimal test channel*. In the hierarchical setup, the conditional distribution is $p(\hat{\mathbf{X}}^1, \hat{\mathbf{X}}^2|\mathbf{X})$, with $\hat{\mathbf{X}}^1$ a coarse reconstruction and $\hat{\mathbf{X}}^2$ a finer one. We aim at both reconstructions lying on the RD curve (optimally trading rate for distortion), with $\hat{\mathbf{X}}^2$ a refinement of $\hat{\mathbf{X}}^1$: $p(\hat{\mathbf{X}}^1, \hat{\mathbf{X}}^2|\mathbf{X}) = p(\hat{\mathbf{X}}^2|\mathbf{X})p(\hat{\mathbf{X}}^1|\hat{\mathbf{X}}^2)$. Such condition is known as *Successive Refinement* (Equitz & Cover, 1991) and is attainable only on some source–distortion pairs (*i.e.* images–MSE).

Theorem 1. (Equitz–Cover characterization (Equitz & Cover, 1991)).
A source–distortion pair is successively refinable from D_1 to D_2 if and only if there exist there exist $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2$ such that $R_1 = R(D_1)$, $R_2 = R(D_2)$ and $\mathbf{X} \to \hat{\mathbf{X}}_2 \to \hat{\mathbf{X}}_1$ is a Markov chain.

Here, the Markov chain implies $\hat{\mathbf{X}}^2$ is a refinement of $\hat{\mathbf{X}}^1$, *i.e.* $p(\hat{\mathbf{X}}^1, \hat{\mathbf{X}}^2|\mathbf{X}) = p(\hat{\mathbf{X}}^2|\mathbf{X})p(\hat{\mathbf{X}}^1|\hat{\mathbf{X}}^2)$. For cumulative rates $0 < R_1 < \cdots < R_T$, the SR notion can be extended at $\{R_k\}_{k=1}^T$ when there exist compatible reconstructions that attain $D(R_k)$ for every k and simultaneously satisfy $\mathbf{X} \to \hat{\mathbf{X}}^{(T)} \to \cdots \to \hat{\mathbf{X}}^{(1)}$; see Appendices B.4 and B.5 for details.

HLVMs and nested codes. A hierarchical latent-variable model (HLVM) (Shu & Ermon, 2022) combines a hierarchical encoder q_ϕ and prior p_θ of learnable parameters ϕ and θ, respectively,

$$q_\phi(\mathbf{Z}_{1:T}|\mathbf{X}) = \prod_{k=1}^{T} q_\phi(\mathbf{Z}_k|\mathbf{X}, \mathbf{Z}_{1:k-1}), \qquad p_\theta(\mathbf{Z}_{1:T}) = \prod_{k=1}^{T} p_\theta(\mathbf{Z}_k|\mathbf{Z}_{1:k-1}), \tag{2}$$

with prefix decoders $\hat{\mathbf{X}}^{(k)} = g_k(\mathbf{Z}_{1:k}; \theta)$. The formulation admits T nested models with prefix $1 \le k \le T$, well defined latents $Z_{1:k}$ and reconstructions $\hat{X}^k$. Such depth-T *nested code* further defines

$$R_k := \frac{I(\mathbf{X}; \mathbf{Z}_{1:k})}{m \log 2}, \qquad D_k := \mathbb{E}\left[d(\mathbf{X}, \hat{\mathbf{X}}^{(k)})\right], \qquad D_k \ge D(R_k), \tag{3}$$

with $R_k - R_{k-1}$ the information added by stage k, and $R_0 := 0$. Both HVAEs and DDPMs are valid HLVMs, with specific characteristics defined as follows.

HVAEs and DDPMs. HVAEs learn the encoder, prior, and decoder by minimizing the negative ELBO. Expressed in bits/dim, with the normalization $1/(m \log 2)$ left implicit,

$$\mathcal{L}_{\text{HVAE}} = \mathbb{E}_{p_{\text{data}}, q_\phi}[-\log p_\theta(\mathbf{X}|\mathbf{Z}_{1:T})] + R_T^{\text{var}}, \quad R_T^{\text{var}} := \mathbb{E}_{p_{\text{data}}} D_{\text{KL}}\big(q_\phi(\mathbf{Z}_{1:T}|\mathbf{X}) \| p_\theta(\mathbf{Z}_{1:T})\big). \tag{4}$$

The variational rate $R_T^{\text{var}} \geq R_T$ is a trainable upper bound, not the true information rate R_T; their exact relationship and the tightness condition are given in Appendix B.8. The ELBO optimizes only the total rate and final reconstruction, leaving intermediate rate allocation and distortion unsupervised.

DDPMs instead prescribe the Gaussian encoder

$$q(\mathbf{Z}_{1:T}|\mathbf{X}) = \prod_{k=1}^{T} q(\mathbf{Z}_k|\mathbf{Z}_{k+1}), \qquad q(\mathbf{Z}_k|\mathbf{Z}_{k+1}) = \mathcal{N}\big(\mathbf{Z}_k; \sqrt{1-\beta_k}\,\mathbf{Z}_{k+1}, \beta_k \mathbf{I}_m\big), \qquad \mathbf{Z}_{T+1} := \mathbf{X}, \tag{5}$$

and supervise every reconstruction with the $\mathbf{X}$-prediction objective (Ho et al., 2020)

$$\mathcal{L}_{\text{DDPM}} = \frac{1}{Tm} \sum_{k=1}^{T} \mathbb{E}_{p_{\text{data}}, q(\mathbf{Z}_k|\mathbf{X})} \left[\|\mathbf{X} - d_k(\mathbf{Z}_k; \theta)\|_2^2 \right], \qquad \hat{\mathbf{X}}^{(k)} := d_k(\mathbf{Z}_k; \theta). \tag{6}$$

The forward Markov property gives $I(\mathbf{X}; \mathbf{Z}_{1:k}) = I(\mathbf{X}; \mathbf{Z}_k)$, so the noise schedule fixes the cumulative rate profile, although the specific mapping from noise level to rate remains source-dependent.

3 Successive-Refinement HVAEs

Theorem 1 describes the ideal progressive representation: after every stage, the reconstruction should use its cumulative information as efficiently as possible, and these reconstructions should be mutually compatible whenever the source permits successive refinement. We now ask how this behavior can be encouraged in a hierarchical generative model. Instead of immediately delivering the objective, we start by analyzing a useful special case: under Gaussian data and MSE, DDPMs already implement the desired successive-refinement channel exactly. We can then use this result to design the principles for a general objective, which we later instantiate in the *Successive-Refinement HVAE* (SR-HVAE).

DDPMs under the successive refinement framework. The following result gives the precise connection between diffusion and the classical communication problem.

Proposition 1. (DDPMs realize an RD-optimal SR channel under isotropic Gaussian/MSE). *Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$ and use MSE distortion. Under idealized population setting, valid noise schedule $(0 < \beta_k < 1)$, per-level positive loss weights, and a shared denoiser class capable of simultaneously representing the MSE-optimal estimate $d_k(\mathbf{Z}_k; \theta^\star) = \mathbb{E}[\mathbf{X} \mid \mathbf{Z}_k]$, $k = 1, \ldots, T$, every global minimizer of Equation* (6) *realizes an RD-optimal successive-refinement autoencoding channel at the cumulative rates induced by the noise schedule.*

The proof is in Appendix C.1; the reasoning follows. The DDPM encoder is additive Gaussian: up to rescaling, the latent at level k is the source plus Gaussian noise of variance s_k^2, larger at low k (coarse levels) and smaller at high k (fine levels). Hence, the noise variance alone fixes how much source information survives, $R_k = \frac{1}{2}\log_2(1 + \sigma^2/s_k^2)$, with $0 < R_1 < \cdots < R_T$. The noise schedule' role is thus of implicit rate targeting: rates are read off the noise levels rather than chosen. On the decoder side, the distortion-minimizing reconstruction at level k is the conditional mean $\mathbb{E}[\mathbf{X}|\mathbf{Z}_k]$, which for a Gaussian source under MSE attains $D_k = D(R_k)$ exactly. The proof further shows that a global minimizer of Equation (6) recovers all D_k at once; since each is an invertible rescaling, the forward Markov chain transfers from latents to reconstructions. Frontier attainment at every level and compatibility across levels are exactly the conditions of Theorem 1.

For a general $\mathbf{X}$ such properties do not hold. Given the latent obtained trough additive gaussian noise, the global minimizer is no longer guaranteed to attain $D(R_k)$ at each level, independently of denoiser's capacity. Furthermore, the optimal denoiser is no longer an invertible rescaling, so the reconstructions do not inherit the Markov chain. Two aspects of DDPM training, however, remain valid: every level receives a precise information budget, and the prefix is trained to minimize distortion w.r.t. the source. What must change is the encoder, which should be able to adapt to the source. But a naively learnable one would lose the benefit of prescribed rate targeting: exactly one failure of vanilla HVAEs, whose ELBO constrains only the total variational rate and the final reconstruction. Together with per-prefix supervision, this sets the principles of our objective: *(i)* prescribe the cumulative rate after each level; *(ii)* supervise the reconstruction given by every prefix.

Principles (i)–(ii): the core objective. The most natural form of learnable encoder yields a top-down HVAE following Equation (2), with prefix reconstructions $\hat{\mathbf{X}}^{(k)} = g_k(\mathbf{Z}_{1:k}; \theta)$. We therefore use it as basis for imposing our principles. Specifically, the idea under principle (i) is to specify a strictly increasing schedule $0 < \gamma_1 < \cdots < \gamma_T$ and penalize the deviation of every cumulative rate from its target. The true rate R_k is defined in Equation (3); HVAE training uses its variational upper bound R_k^{var}, expressed in bits/dim as

$$\begin{aligned} R_k^{\text{var}} &:= \frac{1}{m \log 2} \mathbb{E}_{p_{\text{data}}(\mathbf{X})} \left[D_{\text{KL}}\big(q_\phi(\mathbf{Z}_{1:k}|\mathbf{X}) \| p_\theta(\mathbf{Z}_{1:k})\big) \right] = \sum_{j=1}^{k} \Delta R_j^{\text{var}}, \\ \Delta R_j^{\text{var}} &:= \frac{1}{m \log 2} \mathbb{E}_{p_{\text{data}}(\mathbf{X})} \mathbb{E}_{q_\phi(\mathbf{Z}_{1:j-1}|\mathbf{X})} \left[D_{\text{KL}}\big(q_\phi(\mathbf{Z}_j|\mathbf{X}, \mathbf{Z}_{1:j-1}) \| p_\theta(\mathbf{Z}_j|\mathbf{Z}_{1:j-1})\big) \right]. \end{aligned} \tag{7}$$

Here ΔR_j^{var} is the incremental variational rate of level j. The bound is not tight in general, and the slack has a well-known form (Hoffman & Johnson, 2016):

$$R_k^{\text{var}} = \underbrace{\frac{I_{q_\phi}(\mathbf{X}; \mathbf{Z}_{1:k})}{m \log 2}}_{R_k} + \underbrace{\frac{1}{m \log 2} D_{\text{KL}}\big(q_\phi(\mathbf{Z}_{1:k}) \| p_\theta(\mathbf{Z}_{1:k})\big)}_{\delta_k \geq 0}, \qquad q_\phi(\mathbf{Z}_{1:k}) := \int p_{\text{data}}(\mathbf{x})\, q_\phi(\mathbf{Z}_{1:k}|\mathbf{x})\, d\mathbf{x}. \tag{8}$$

Its first term is the information the prefix actually carries about the source, the quantity successive refinement constrains. Its second term, which we call the *prior slack* δ_k, is the divergence between the *aggregate posterior* $q_\phi(\mathbf{Z}_{1:k})$, and the prior the generative path samples from (Rosca et al., 2018; Tomczak & Welling, 2018; Dai & Wipf, 2019). Only R_k^{var} is computable during training, and δ_k inflates it: matching γ_k in variational terms guarantees that the prefix carries *at most* γ_k bits of source information, with equality only when $\delta_k = 0$. Under *variational tightness*, however, matching the rate targets gives every level a nonzero increment in true source information.

Principle (ii) supervises instead the reconstruction available from each prefix. Using MSE/dim,

$$D_k := \frac{1}{m} \mathbb{E}_{p_{\text{data}}(\mathbf{X})} \mathbb{E}_{q_\phi(\mathbf{Z}_{1:k}|\mathbf{X})} \left[\|\mathbf{X} - \hat{\mathbf{X}}^{(k)}\|_2^2 \right]. \tag{9}$$

Taken together at an arbitrary level k, the two principles suggest the objective $|R_k^{\text{var}} - \gamma_k| + D_k$: the first term keeps the cumulative rate near its target, while the second encourages the best reconstruction at that rate. Being such objective a sum of the two terms, however, brings two additional considerations. First, increasing rate automatically enables a lower distortion. Second, the balance depends on the slope of the RD curve at the given rate point. Intuitively, an optimization process is willing to pay one additional bit of rate at coarse levels, where the corresponding distortion gain is larger. Specifically, given a reference curve $D_{\text{ref}}(R)$, decreasing the rate from γ_k by a small amount $\delta > 0$ increases the minimum reference distortion as $D_{\text{ref}}(\gamma_k - \delta) - D_{\text{ref}}(\gamma_k) \approx |D'_{\text{ref}}(\gamma_k)|\delta$. Thus, the measure $s_k := |D'_{\text{ref}}(\gamma_k)|$ converts a rate displacement at level k into its local distortion-equivalent cost. The balanced per-level objective is therefore $\lambda s_k |R_k^{\text{var}} - \gamma_k| + D_k$, where $\lambda > 0$ controls the exchange between rate enforcement and distortion reduction. Such objective balances the rate and distortion terms at a single k. However, the scales s_k may themselves vary substantially across the hierarchy, causing coarse levels (where the slope is steep) to dominate the sum. We therefore normalize the entire level-k objective by $1/s_k$, allowing to express every level in the same local rate-equivalent units. Combining everything gives the final objective

$$\mathcal{L}_{\text{SR-HVAE}}^{(1,2)} = \frac{1}{T} \sum_{k=1}^{T} \left[\underbrace{\lambda\, |R_k^{\text{var}} - \gamma_k|}_{\text{principle (i)}} + \underbrace{\frac{D_k}{s_k}}_{\text{principle (ii)}} \right], \qquad s_k := |D'_{\text{ref}}(\gamma_k)|\,. \tag{10}$$

In practice, the reference curve $D_{\text{ref}}(R)$ is source-specific, and we discuss its choice in Section 4. After normalization, λ has a direct exchange-rate interpretation: a deviation of one bit/dim from the target costs λ rate-equivalent units and must therefore produce more than λ units of normalized distortion reduction to be favored. In particular, $\lambda = 1$ matches the local reference trade-off, whereas $\lambda > 1$ prioritizes rate enforcement.

Formal objective guarantees. Equation (10) offers a balanced objective for optimizing the two proposed principles. We now show that it recovers a *valid* schedule and formalize the distortion trade-off it selects at the achieved rates. A schedule is valid for a model family when it is strictly increasing and the family contains a nested code attaining all its targets with finite cumulative distortion.

Proposition 2. (Optimality at the achieved rates).
Let $\{\gamma_k\}_{k=1}^{T}$ be a valid rate schedule. Under the population-level, global-optimization, and variational-tightness assumptions detailed in Appendix C.2, every global minimizer of Equation (10) *has aggregate absolute rate error $O(\lambda^{-1})$ as $\lambda \to \infty$. At its achieved information-rate allocation, it minimizes the slope-normalized cumulative distortion $\sum_{k=1}^{T} D_k/s_k$ over all model-realizable nested codes with that allocation and is therefore Pareto-optimal within this fixed-rate comparison class.*

Thus, the first term is a penalty relaxation of Principle (i), whereas the second term selects a minimum slope-normalized cumulative-distortion solution once the rates are fixed, compatible with Principle (ii). The following consequence identifies when such objective reaches the Rate–Distortion frontier.

Corollary 2.1. (SR / non-SR dichotomy).
Under the conditions of Proposition 2, consider an ordered allocation achieved by a global minimizer. If the source–distortion pair is successively refinable at this allocation and the model class contains a corresponding RD-optimal nested code, then every global minimizer with that allocation attains $D_k = D(R_k)$ at every level. If the pair is not successively refinable at the achieved allocation, every nested code with that allocation has a strict RD gap at one or more levels.

The results state that Equation (10) correctly implement the principles whether or not the source is SR. In the SR case, it recovers compatible frontier-optimal prefixes whenever the corresponding code is model-realizable. Otherwise, it selects a Pareto-optimal nested compromise, while at least one RD gap is unavoidable. Proofs are given in Appendices C.2 and C.3.

Learnable *vs*. prescribed encoders. Equation (10) optimizes encoder and decoder jointly, so an HLVM trained with it is truly hierarchical in the sense our paper asks for: every latent adds new information, and that information makes the resulting prefix a better reconstruction. A DDPM is the special case in which the first term vanishes by construction: the prescribed noising process fixes R_k at the rates its schedule induces, leaving the per-prefix distortions term alone (Equation (6)). Satisfying principle (i) by construction rather than by optimization carries the price identified above: when the source departs from the one the channel was designed for, no decoder capacity places its prefixes back on the RD frontier. Nonetheless, the same rigidity buys something a learnable encoder must earn. Because the encoder is prescribed, R_k is a functional of the source and the noise schedule alone: it needs no variational bound, and Equation (8) does not apply. A learnable encoder has no such guarantee, its aggregate posterior moves during training and the prior must track it, inflating the rate the objective controls and degrading the samples drawn from p_θ. Neither design is therefore perfect, but only one of the two failures is recoverable: a misspecified fixed encoder cannot be repaired, whereas a loose bound can be closed by acting on the prior. We therefore argue the correct choice is a learnable encoder whose bound is tightened, for instance by refitting the prior to the aggregate posterior it induces. Section 4 measures δ_k, isolates its generative cost by decoding from the aggregate posterior rather than the prior, and verifies refitting the prior closes part of the gap.

4 Experiments

We evaluate our objective in two regimes. On analytic sources, where the rate–distortion structure is known and we can verify the successive refinement predictions directly. On images, where we test how a meaningful hierarchy acts on real data, and weather it improves generative metrics. Full experimental details, protocols and ablations are in Appendices D and E.

4.1 Analytic sources

Setup. We use two 64-dimensional i.i.d. sources under MSE. The first is an isotropic Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I}_{64})$, which is SR. The second is a symmetric binary Gaussian mixture (SBGM, μ=2, σ=1), which Chow & Berger (1997) show is *not* refinable in a characterised range, with critical rate $R_{\text{crit}} = 0.913$ bits/dim. Since $D(R)$ is known for both, we use it as the reference curve D_{ref} of

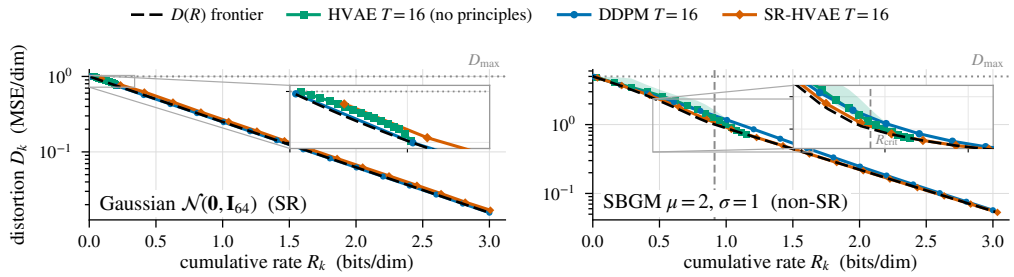

Figure 1: Rate–distortion plane for the analytic sources. Each curve traces the median (R_k, D_k) pairs over seeds against the source frontier $D(R)$ (dashed). **(a)** On the Gaussian source the DDPM and SR-HVAE both track the frontier over the whole range, while the HVAE stalls at $R_T = 0.21$. **(b)** On SBGM the HVAE stalls at $R_T = 1.15$, collapsing in 5 of 10 seeds (excluded from the curve). Both SR-HVAE and DDPM depart from the frontier in the neighbourhood of R_{crit}, exactly as the theory predicts, but the DDPM's prescribed encoder pays a substantially larger rate–distortion cost.

Table 1: Analytic sources generative metrics, mean over 10 seeds. δ_T reports a bracket estimate of the prior slack on our SR-HVAE, in percent of R_T^{var}. SW2 is reported over axis-aligned projections (ax). Best among T=16 models in **bold**, best overall underlined.

Model	*Gaussian* $\mathcal{N}(\mathbf{0}, \mathbf{I}_{64})$ *— SR*			*SBGM* μ=2, σ=1 *— non-SR*		
	$\delta_T \downarrow$	ax-SW2 $\downarrow$	MMD $\downarrow$	$\delta_T \downarrow$	ax-SW2 $\downarrow$	MMD $\downarrow$
DDPM T=1000	–	0.0157	$3.2{\cdot}10^{-5}$	–	0.5144	$4.3{\cdot}10^{-5}$
DDPM T=16	–	0.1349	$6.6{\cdot}10^{-3}$	–	0.5698	$7.0{\cdot}10^{-3}$
HVAE T=16	–	0.5069	$7.9{\cdot}10^{-2}$	–	1.3372	$8.0{\cdot}10^{-2}$
SR-HVAE T=16 (ours)	[0.30, 1.28]	**0.0210**	$\mathbf{1.1{\cdot}10^{-4}}$	[0.00, 0.18]	**0.0372**	$\mathbf{2.9{\cdot}10^{-5}}$

Equation (10), and can exactly measure the distance from the frontier. All models use T=16 levels over the cumulative rate range $[0.01, 3.0]$ bits/dim. We compare four arms: a vanilla HVAE, SR-HVAE, a T=16 DDPM, and a T=1000 DDPM. The rate range is enforced exactly for SR-HVAE and, after noise-schedule calibration, for both DDPMs; the HVAE is left unconstrained, since the ELBO cannot express a rate target. All runs use 10 seeds and early stopping. We keep $\lambda = 1$ throughout, which matches the local reference trade-off, and study the stronger rate enforcement obtained with larger λ in Appendix D.8. Figure 1 locates every model on the rate–distortion plane against the frontier. Its x-axis is the true information rate for the DDPMs, whereas for the HVAE and SR-HVAE it is the variational upper bound of Equation (8). We therefore bracket the prior slack δ_T between estimated bounds on the true rate and report it in Table 1 as a percentage of R_T^{var}, which quantifies that displacement. The table also reports the distributional distance between model samples and the source, through sliced Wasserstein-2 on the 64 coordinate axes (ax-SW2), and through Gaussian-kernel MMD. Further details on experimental setup, implementation, and results in Appendix D.

Results and discussion. Figure 1a confirms Proposition 1 directly: on the Gaussian source the T=16 DDPM attains $D_k = D(R_k)$ at every level. Our objective reproduces that behavior with a learnable encoder of identical capacity to the vanilla HVAE, up to a small offset due to finite training and to the variational upper bound, since the true rate lies between 0.30% and 1.28% below the plotted value (Table 1). Corollary 2.1 predicts that on SBGM no nested code can be frontier-optimal at every level. In Figure 1b both families depart, in the neighbourhood of the critical rate the theory names (Chow & Berger, 1997). The size of the departure separates the two designs: the learnable encoder has lower peak and returns to the frontier immediately afterwards, whereas the prescribed Gaussian channel never rejoins it within the range. This shows the cost of the fixed encoder on a non-gaussian source: the DDPM hierarchy remains valid, but it cannot be rate–distortion optimal. The vanilla HVAE fails on both sources as its ELBO constrains only the total variational rate and the final reconstruction. Table 1 shows the hierarchy effect on generative quality. On the Gaussian source the T=1000 DDPM is best on every metric, as Proposition 1 leads one to expect. On SBGM SR-HVAE surpasses the T=1000 DDPM on every generative metric, as the variational rate sits close to the true rate (0.18% slack at most).

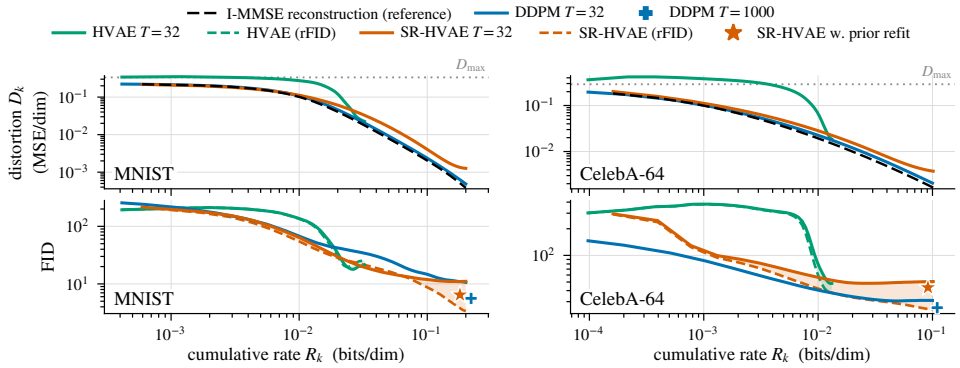

Figure 2: **Top:** distortion–rate plots for all models against the IMMSE reference, SR-HVAE and DDPM at extended rate budget. **Bottom:** FID–rate plots for the same models, with dash lines denoting rFID. Endpoint markers are the final FID after prior refit and a 1000-step DDPM.

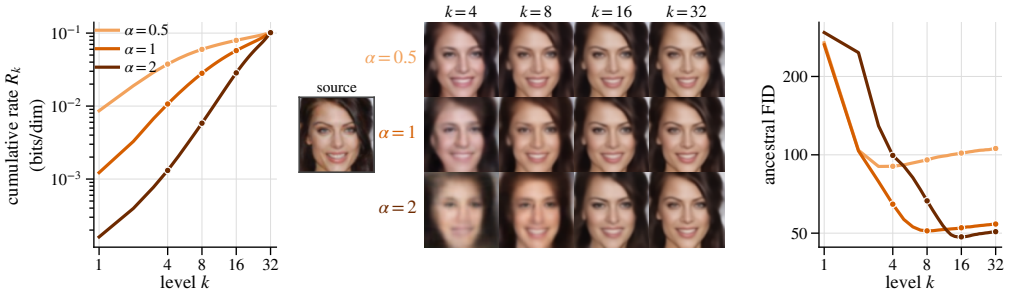

Figure 3: Left: the same SR-HVAE model is trained with three different rate allocations, governed by $\alpha = \{0.5, 1.0, 2.0\}$. Center: allocations decide how the hierarchy builds samples. $\alpha = 0.5$ spends most latents in details, $\alpha = 2$ builds samples more progressively. Right: allocations have direct effects on FID. More latents at coarse rate improves generative quality.

4.2 Image sources

Setup. We train every model for $T = 32$ levels on MNIST and CelebA-64, under two rate budgets. The *matched* budget sets γ_T to the total rate the vanilla HVAE reaches on its own (0.0325 and 0.0132 bits/dim), so that model families are compared at the allocation the ELBO selects; the *extended* budget raises it to 0.2 and 0.1 bits/dim, respectively. The reference curve D_{ref} is obtained from the I-MMSE relation (Guo et al., 2005) on an independent, larger diffusion model, representing an *achievable* curve rather than the unknown $D(R)$. DDPMs use the paired noise schedule obtained by inverting the same reference table. We measure generative quality in terms of FID (Heusel et al., 2017), and improved precision and recall (Kynkäänniemi et al., 2019). On learnable encoder models, we draw samples both from the *ancestral* generative prior and the *reconstruction* posterior. Since the two share the same decoder and rate allocation, the distance between them isolates the effect of the prior slack δ_k from the quality of the hierarchy.

Results and discussion. Figure 2 tracks, level by level, the distortion and FID metrics as rate increases, showing how each model builds its own hierarchy. HVAE models (learnable encoder) include the rFID result, tracking the differences between sampling from priors *vs*. posteriors (*i.e.* the δ_k slack). Table 2 collects the final generative metrics, including the matched rate models (second and third row) and precision/recall columns. Read together, the results show how the ELBO objective settles the model at a low total rate (0.032 and 0.013 bits/dim), causing a first generative gap. Increasing the rate region of both SR-HVAE and DDPM (see Table 2) immediately improves all generative metrics. On the RD–plane (Figure 2 (top), HVAE also fails at building a meaningful hierarchy. Without per-prefix distortion supervision, 4 of 32 prefixes on MNIST and 12 of 32 on CelebA reconstruct no better than the data mean, while both DDPM and SR-HVAE construct samples progressively. We further find the allocation *within* a fixed range matters as well (Figure 3). The exponent α sets it, with α=0.5 reserving more latents for high rates and α=2 the opposite. Placing them at coarse levels, where the RD slope is steeper, lets SR-HVAE's build samples progressively and pays in FID: moving from α=0.5 to α=2 gives 26.6 to 11.1 FID improvement on MNIST and 105.5

Table 2: **top:** generative metrics for each model, sampled from prior. **bottom** same metrics with samples from the reconstruction arm, showing the theoretical generative lower bound on a model with tight variational rate (no prior slack). **Bold** denotes best result per block, *underline* the second-best.

		MNIST				CelebA-64			
Model	T	R_T	FID ↓	prec. ↑	rec. ↑	R_T	FID ↓	prec. ↑	rec. ↑
HVAE	32	0.032	**21.59**	*0.237*	**0.288**	0.013	*48.26*	0.286	*0.005*
SR-HVAE (matched rate)	32	0.033	*22.80*	**0.294**	*0.236*	0.014	53.14	*0.309*	0.004
DDPM (matched rate)	32	0.032	46.75	0.039	0.114	0.013	**38.37**	**0.436**	**0.028**
SR-HVAE	32	0.200	11.09	0.379	0.632	0.101	50.79	0.367	0.024
SR-HVAE (prior refit)	32	0.200	*6.46*	**0.592**	0.645	0.101	*43.32*	0.435	0.024
DDPM	32	0.200	10.47	0.539	*0.662*	0.100	*30.84*	**0.577**	*0.118*
DDPM	1000	0.200	**5.61**	*0.548*	**0.670**	0.100	**25.64**	*0.572*	**0.198**
reconstruction arm (prior-free upper bound)									
HVAE	32	0.032	*26.03*	*0.416*	*0.171*	0.013	*40.83*	*0.370*	*0.013*
SR-HVAE	32	0.200	**3.36**	**0.968**	**0.942**	0.101	**24.21**	**0.565**	**0.225**

to 50.8 on CelebA (Figure 3 right). We include a torought ablation on both datasets in Appendix E.7. All results point towards the same direction, which is the main message of our paper: *rate control and prefix-reconstruction supervision actively help HLVMs in building a meaningful hierarchy, and this directly improves generative metrics*.

Given the common objective, what distinguishes HLVM families is the architecture. A prescribed encoder cannot adapt to the source; a learned one can, but pays the prior–posterior mismatch. Table 2 and Figure 2 suggest which price is worth paying: the SR-HVAE reconstruction arm shows the achievable metrics with tight bound, improving over DDPM's fixed encoder overall. Hence, a fully learnable hierarchy can obtain the best results with a low δ_k slack. Refitting the prior alone recovers part of this slack, empirically confirming this is the right direction to pursue.

5 Discussion and Conclusions

Limitations and future work. Image experiments are limited by our computational bounds. They use a single seed, low-resolution datasets and low model capacity. The main claims remain valid, as the focus is on how a hierarchy spends a given budget, not on absolute sample quality. Capacity limits bound especially SR-HVAEs, which carry the challenging objective of multi-step encoder-decoder optimization. Hence, at 64×64 the prescribed encoder still wins outright: refitting recovers the gap partially, leaving SR-HVAE at 43.3 against 30.8 for DDPM on the ancestral arm. Two quantities are weaker than we would like: D_{ref} must be trained before any schedule exists when $D(R)$ is unknown, and δ_k is bracketed on analytic sources and projected on images, rather than exact. Scaling the same objective to larger models and sources, closing the prior slack with more expressive hierarchical priors, and replacing D_{ref} with a tighter or learned reference are the natural next steps.

Conclusions. The final message of this paper is that generative hierarchies should not be evaluated only at their final output. If a model claims to generate progressively, then each intermediate stage should have a measurable role, actively contributing with new information and producing a meaningful reconstruction. We identified in successive refinement the correct theory to deal with these aspects, deriving two principles (control the cumulative rate at each stage, and supervise every prefix reconstruction) and synthesizing them into a common objective for hierarchical models. The same view reads existing models. DDPMs are the special case in which the encoder is prescribed, satisfying the first principle by construction at the cost of rate–distortion optimality away from the Gaussian source; the HVAE ELBO satisfies neither. Experiments confirm both readings. On analytic sources our objective matches SR predictions, both on refinable and non-refinable sources. On images it constructs better hierarchy and enables greater flexibility: deciding which rate region to target and how to spread latents across it are choices that directly improve generative quality. We further find that the same objective can be pursued by different encoder instances. Decoded from its own posterior, SR-HVAE surpasses an equal-depth DDPM on both datasets, so the learned hierarchy is the better one, as theory predicts. In practice, prior-posterior mismatch can reduce generative performance, and refitting the prior is the immediate solution that preserves the hierarchy while reducing the gap.

AI USE STATEMENT

In this work, we used generative AI tools to help interpret and discuss results, to provide feedback on research methodology and experiments, and to assist in the writing of proofs: an LLM assisted in reading experimental outputs, in checking which of our claims the measurements actually supported, in proposing control experiments and evaluation protocols, which we then specified and ran ourselves, and in rendering our proof sketches into formal prose and checking the resulting arguments for correctness. We have not used generative AI tools to generate synthetic data sets, to develop the theoretical model or conceptual framework, to formulate mathematical claims, to provide critical ingredients for proving them, or to propose or refine hypotheses: the successive-refinement framework, the objective of Equation (10), and the statements of Proposition 1, Proposition 2 and Corollary 2.1, together with the arguments establishing them, are our own; AI assistance on the proofs was limited to formalizing and verifying arguments we had already sketched. Assistance with translation, dataset cleaning and reformatting, and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to draft parts of the paper and its appendices, to edit the manuscript for readability and length, to suggest its structure, to create the scripts that render Figure 1, Figure 2 and Figure 3, to write and edit analysis and evaluation code, to identify relevant literature, and to format references. We have reviewed all AI-assisted work. Every AI-drafted or AI-edited passage was read by an author and verified against our experimental records before inclusion, and every numerical value appearing in the text, tables and figures was checked against the stored result files. Each reference found with AI assistance was retrieved and read in its original form, and its bibliographic entry checked against the publisher record. AI-written analysis and figure code was reviewed by an author and its outputs cross-checked against independently computed values. To the best of our knowledge, the claims, results and code reported here are original and correct. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions behind each theoretical claim are stated with the claim in Section 3 and discharged in full in Appendix C, with the conventions they rely on collected in Appendix B. On the experimental side, Appendix D and Appendix E give the complete protocol for the analytic and the image sources respectively: source definitions and preprocessing, architectures and parameter counts, the rate schedules together with the $r_{\min}$ and $r_{\max}$ used by every arm, and optimizer settings and training iterations. Because every image schedule is derived from the I-MMSE reference rather than from a known $D(R)$, Appendix E.6 documents separately how D_{ref} is constructed, which independent denoisers it is computed from, and the numerical validation of the quadrature. Code,trained models and the complete result files will be released upon acceptance.

REFERENCES

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken ELBO. In *International Conference on Machine Learning*, 2018.

Toby Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. In *International Conference on Learning Representations*, 2017.

Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2020.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

James Chow and Toby Berger. Failure of successive refinement for symmetric gaussian mixtures. *IEEE Transactions on Information Theory*, 43(1):350–352, 1997.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006. ISBN 0471241954.

Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019.

William HR Equitz and Thomas M Cover. Successive refinement of information. *IEEE Transactions on Information Theory*, 37(2):269–275, 1991.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *NeurIPS Workshop on Advances in Approximate Bayesian Inference*, 2016.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in Neural Information Processing Systems*, 2021.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Tamas Linder and Ram Zamir. On the asymptotic tightness of the shannon lower bound. *IEEE Transactions on Information Theory*, 40(6):2026–2031, 1994.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021.

Beatrix Miranda Ginn Nielsen, Anders Christensen, Andrea Dittadi, and Ole Winther. DiffEnc: Variational diffusion with a learned encoder. In *International Conference on Learning Representations*, 2024.

Jan Østergaard and Ram Zamir. Incremental refinement using a gaussian test channel. In *IEEE International Symposium on Information Theory Proceedings*, pp. 2233–2237. IEEE, 2011.

Bixio Rimoldi. Successive refinement of information: Characterization of the achievable rates. *IEEE Transactions on Information Theory*, 40(1):253–259, 1994.

Mihaela Rosca, Balaji Lakshminarayanan, and Shakir Mohamed. Distribution matching in variational inference. *arXiv preprint*, arXiv:1802.06847, 2018.

Kenneth Rose. A mapping approach to rate-distortion computation and analysis. *IEEE Transactions on Information Theory*, 40(6):1939–1952, 1994.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Claude Elwood Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record*, 1959.

Rui Shu and Stefano Ermon. Bit prioritization in variational autoencoders via progressive coding. In *International Conference on Machine Learning*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in Neural Information Processing Systems*, 2016.

Lucas Theis, Tim Salimans, Matthew D Hoffman, and Fabian Mentzer. Lossy compression with gaussian diffusion. *arXiv preprint*, arXiv:2206.08889, 2022.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, 2024.

Jakub Tomczak and Max Welling. VAE with a VampPrior. *arXiv preprint*, arXiv:1705.07120, 2018.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems*, 2020.

Yibo Yang, Justus Will, and Stephan Mandt. Progressive compression with universally quantized diffusion models. In *International Conference on Learning Representations*, 2025.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

A Related Works

Hierarchical generation and rate–distortion readings of the ELBO. Deep HVAEs (Child, 2020; Vahdat & Kautz, 2020) with top-down inference (Sønderby et al., 2016) achieve strong likelihoods, but nothing in the ELBO requires their latents to be used: with a sufficiently expressive decoder the model prefers to ignore the code entirely (Chen et al., 2017; Zhao et al., 2019). Alemi et al. (2018) give the sharpest statement of why. Deriving variational bounds on $I(\mathbf{X};\mathbf{Z})$ and the induced rate–distortion curve, they exhibit a *family of models with identical ELBO* occupying very different points of that plane, so the objective does not determine where a model sits, and they make the operating point an explicit choice rather than a by-product. Their analysis concerns a single latent and a single point. Our principle (i) is the hierarchical counterpart: an HLVM has one such point per prefix, and leaving all T of them unspecified is what Section 3 shows the ELBO does.

Noise schedules and where rate is spent. Kingma et al. (2021) show that the continuous-time diffusion ELBO depends on the noise schedule only through the signal-to-noise ratio at its endpoints, and is otherwise *invariant to the schedule's shape*. That invariance is a statement about the continuous limit, and our finite-T measurements are its counterpart: at $T{=}32$ the exponent α moves FID by $8.5\times$ on MNIST, at $T{=}1000$ by only $1.5\times$ (Appendix E.7). The schedule matters exactly when steps are scarce, which is the regime in which a hierarchy is a hierarchy rather than a discretisation. Related efforts reshape the same budget without measuring it on the RD plane: Choi et al. (2022) reweight the per-step loss to prioritise the noise levels carrying perceptually salient content, and Nielsen et al. (2024) add source adaptivity through a data- and depth-dependent mean function in the diffusion process, retaining a per-step objective and leaving per-level RD optimality unexamined.

Diffusion models as lossy compressors. Ho et al. (2020) observe that a DDPM admits a progressive lossy code and report the resulting rate–distortion points. Theis et al. (2022) turn this into a practical codec through reverse channel coding, and Yang et al. (2025) extend it with universal quantization to a single progressive codec competitive across bitrates. These treat the model as one compressor and ask how well it codes overall, not whether each intermediate level is individually RD-optimal.

The closest predecessor is Shu & Ermon (2022), who adopt the progressive coding hypothesis, *i.e*. that HLVMs good at progressive lossy compression generate well, and show empirically that conventionally trained HVAEs are poor progressive coders, a finding our Figure 2 corroborates from the rate–distortion side. Their remedy partitions the hierarchy at a single prescribed cut K into $(\mathbf{Z}_{1:K}, \mathbf{Z}_{K+1:T})$, minimises the rate of the first group subject to a distortion constraint, then freezes it and trains the remainder by ELBO. They note the procedure extends to more groups but implement the binary split. Our objective instead constrains every prefix at once, in a single stage, and because it is derived from successive refinement rather than posited, it also says *when* such a hierarchy can be simultaneously optimal and when, as on Corollary 2.1's non-refinable side, no allocation makes it so.

Successive refinement. SR theory (Equitz & Cover, 1991; Rimoldi, 1994) characterises when nested lossy codes can be RD-optimal at every prefix simultaneously, the Markov condition of Theorem 1 supplying the characterisation we build on. Chow & Berger (1997) exhibit a source–distortion pair where this fails, which we use as the non-refinable probe of Section 4.1, and Østergaard & Zamir (2011) quantify the incremental-refinement cost for Gaussian test channels, the setting in which Proposition 1 places DDPMs. To our knowledge, SR has not previously been used as a design framework for hierarchical generative models: we take the step from characterising when such hierarchies *can* exist to an objective that trains them.

B Extended Background

This appendix provides the information-theoretic background underlying Section 2. We first review the information measures and operational source-coding setting used by rate–distortion (RD) theory. We then state the two-stage and multistage successive-refinement (SR) problems, work through the Gaussian/MSE example, discuss the non-SR symmetric binary Gaussian mixture, and finally connect the classical quantities to HVAEs and DDPMs. Comprehensive introductions are given by Shannon (1948; 1959) and Cover & Thomas (2006), with a dedicated treatment of RD theory in Berger (1971).

B.1 Information measures and conventions

Entropy and differential entropy. For a discrete random variable $\mathbf{X}$ with probability mass function $p(\mathbf{x})$, its entropy is

$$H(\mathbf{X}) := -\sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x}). \tag{11}$$

Entropy measures the average uncertainty of the source. Shannon's lossless source-coding theorem states that, for an i.i.d. source and asymptotically long blocks, the expected number of bits required per source symbol can approach $H(\mathbf{X})$ but cannot be made smaller without incurring decoding errors (Shannon, 1948; Cover & Thomas, 2006).

For a continuous random variable with density $p(\mathbf{x})$, the analogous quantity is the differential entropy

$$h(\mathbf{X}) := -\int_{\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x}. \tag{12}$$

Unlike discrete entropy, differential entropy may be negative and depends on the coordinate system. It should therefore not be interpreted directly as a code length. The KL divergence and mutual information defined below remain non-negative for both discrete and continuous variables whenever they are well-defined.

KL divergence and mutual information. For distributions p and q on a common space, the KL divergence is

$$D_{\mathrm{KL}}(p\|q) := \mathbb{E}_p\left[\log \frac{p}{q}\right] \geq 0, \tag{13}$$

with equality iff $p = q$ almost everywhere. The mutual information between $\mathbf{X}$ and $\mathbf{Y}$ can be written as

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &:= D_{\mathrm{KL}}\big(p(\mathbf{X},\mathbf{Y})\|p(\mathbf{X})p(\mathbf{Y})\big) \\ &= \mathbb{E}_{p(\mathbf{X})}\left[D_{\mathrm{KL}}\big(p(\mathbf{Y}|\mathbf{X})\|p(\mathbf{Y})\big)\right]. \end{aligned} \tag{14}$$

Thus $I(\mathbf{X};\mathbf{Y})$ is the average amount by which observing $\mathbf{X}$ changes the distribution of $\mathbf{Y}$. Conditional mutual information and the chain rule satisfy

$$I(\mathbf{X};\mathbf{Y},\mathbf{Z}) = I(\mathbf{X};\mathbf{Y}) + I(\mathbf{X};\mathbf{Z}|\mathbf{Y}). \tag{15}$$

Markov chains and data processing. The notation $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ means that $\mathbf{X}$ and $\mathbf{Z}$ are conditionally independent given $\mathbf{Y}$. Equivalently,

$$p(\mathbf{z}|\mathbf{x},\mathbf{y}) = p(\mathbf{z}|\mathbf{y}) \quad \Longleftrightarrow \quad I(\mathbf{X};\mathbf{Z}|\mathbf{Y}) = 0. \tag{16}$$

Every Markov chain obeys the data-processing inequality

$$\mathbf{X} \to \mathbf{Y} \to \mathbf{Z} \quad \Longrightarrow \quad I(\mathbf{X};\mathbf{Z}) \le I(\mathbf{X};\mathbf{Y}). \tag{17}$$

Processing a representation cannot increase the information it contains about the source. This inequality supplies the key bridge from latent-variable models to RD lower bounds in Appendix B.8.

Units. The logarithm base determines the unit: base-2 logarithms give bits and natural logarithms give nats. One nat equals $1/\log 2$ bits. For a vector source $\mathbf{X} \in \mathbb{R}^m$, this work reports information quantities in bits per dimension by dividing a KL divergence computed in nats by $m \log 2$. Distortion is reported as per-dimension MSE,

$$d(\mathbf{x},\hat{\mathbf{x}}) := \frac{1}{m}\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2. \tag{18}$$

B.2 Operational lossy source coding

Source blocks and codes. Let $\mathbf{X}^n = (\mathbf{X}_1,\dots,\mathbf{X}_n)$ be a block of n i.i.d. source symbols drawn from $p(\mathbf{x})$. A fixed-length lossy source code consists of an encoder and decoder

$$f_n : \mathcal{X}^n \to \{1,\dots,M_n\}, \qquad g_n : \{1,\dots,M_n\} \to \hat{\mathcal{X}}^n. \tag{19}$$

The encoder transmits an index $f_n(\mathbf{X}^n)$, from which the decoder constructs $\hat{\mathbf{X}}^n = g_n(f_n(\mathbf{X}^n))$. The rate and average additive distortion are

$$R_n := \frac{1}{n}\log M_n, \qquad D_n := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[d(\mathbf{X}_i,\hat{\mathbf{X}}_i)\right]. \tag{20}$$

The rate is measured per source symbol; an additional division by m gives bits per source dimension for m-dimensional symbols.

Definition 1. (Achievable rate–distortion pair). *A pair (R, D) is achievable if, for every $\epsilon > 0$ and all sufficiently large n, there exists a code satisfying $R_n \le R + \epsilon$ and $D_n \le D + \epsilon$. The operational rate–distortion function is the infimum of achievable rates at distortion D.*

The asymptotic blocklength is essential. A scalar or finite-block quantizer can fail to attain the information-theoretic frontier even when the source is successively refinable asymptotically. In particular, optimal finite scalar quantizers need not form nested partitions.

B.3 Rate–distortion theory

Information characterization. For a memoryless source under an additive distortion measure, Shannon's lossy source-coding theorem equates the operational rate–distortion function with a single-letter mutual-information optimization (Shannon, 1959; Cover & Thomas, 2006; Berger, 1971).

Theorem 2. (Rate–distortion theorem). *Under the standard regularity conditions for the source and distortion alphabets,*

$$R(D) = \inf_{\substack{p(\hat{\mathbf{x}}|\mathbf{x}):\\ \mathbb{E}[d(\mathbf{X},\hat{\mathbf{X}})]\le D}} I(\mathbf{X};\hat{\mathbf{X}}). \tag{21}$$

Equivalently, the minimum distortion permitted by information rate R is

$$D(R) = \inf_{\substack{p(\hat{\mathbf{x}}|\mathbf{x}):\\ I(\mathbf{X};\hat{\mathbf{X}})\le R}} \mathbb{E}\left[d(\mathbf{X},\hat{\mathbf{X}})\right]. \tag{22}$$

The conditional distribution $p(\hat{\mathbf{x}}|\mathbf{x})$ is called a *test channel*. It is a probabilistic description of an asymptotically optimal encoder–reconstruction pair, not necessarily a physical noise channel used by an implementation. A test channel attaining Equation (21) is RD-optimal at distortion D.

Achievability and converse. The achievability part of the theorem constructs long random codebooks whose reproduction sequences behave according to an optimizing test channel. The converse states that no code operating below $R(D)$ can attain distortion D. Consequently, every feasible operating point lies on or above the distortion–rate curve:

$$R \geq R(D) \quad \Longleftrightarrow \quad D \geq D(R), \tag{23}$$

where $D(R)$ is understood as the distortion–rate function rather than necessarily a strict algebraic inverse.

Basic properties. The function $R(D)$ is non-increasing and convex in D. The zero-rate distortion is

$$D_{\max} := \inf_{\hat{\mathbf{x}} \in \hat{\mathcal{X}}} \mathbb{E}\left[d(\mathbf{X}, \hat{\mathbf{x}})\right], \qquad R(D) = 0 \quad \text{for } D \geq D_{\max}. \tag{24}$$

For squared error, $D_{\max}$ is the variance when the reproduction is the source mean. Interior points of the RD frontier can be obtained from the Lagrangian problem

$$\inf_{p(\hat{\mathbf{x}}|\mathbf{x})} \left\{ I(\mathbf{X}; \hat{\mathbf{X}}) + \lambda \mathbb{E}\left[d(\mathbf{X}, \hat{\mathbf{X}})\right] \right\}, \qquad \lambda \geq 0. \tag{25}$$

Sweeping λ traces supporting points of the convex RD curve. This is the classical counterpart of rate–distortion Lagrangians used in learned compression and latent-variable models.

The Shannon lower bound. For a scalar continuous source under squared-error distortion, the Shannon lower bound is

$$R(D) \geq h(\mathbf{X}) - \frac{1}{2} \log_2(2\pi e D). \tag{26}$$

Equality holds when there is an RD-optimal backward channel of the form $\mathbf{X} = \hat{\mathbf{X}} + \mathbf{N}$, where $\mathbf{N} \sim \mathcal{N}(0, D)$ is independent of $\hat{\mathbf{X}}$ and the induced reproduction distribution is valid. The bound is asymptotically tight under broad conditions as $D \to 0$ (Linder & Zamir, 1994).

Latent representations obey the RD lower bound. Suppose an encoder produces a representation $\mathbf{Z}$ and a decoder forms $\hat{\mathbf{X}} = g(\mathbf{Z})$. Then $\mathbf{X} \to \mathbf{Z} \to \hat{\mathbf{X}}$, and Equation (17) gives

$$I(\mathbf{X}; \hat{\mathbf{X}}) \leq I(\mathbf{X}; \mathbf{Z}) =: R \quad \Longrightarrow \quad \mathbb{E}\left[d(\mathbf{X}, \hat{\mathbf{X}})\right] \geq D(R). \tag{27}$$

This is why a latent model can be evaluated by plotting its achieved information rate and reconstruction distortion against the source RD curve.

B.4 Two-stage successive refinement

Operational problem. Successive refinement asks whether a lossy description designed for a coarse distortion can remain useful when additional information is later transmitted (Equitz & Cover, 1991). For a block $\mathbf{X}^n$, consider base and refinement encoders

$$\begin{aligned} f_{1,n} &: \mathcal{X}^n \to \mathcal{M}_{1,n}, \\ f_{2,n} &: \mathcal{X}^n \to \mathcal{M}_{2,n}, \end{aligned} \tag{28}$$

with decoders

$$\begin{aligned} g_{1,n} &: \mathcal{M}_{1,n} \to \hat{\mathcal{X}}^n, \\ g_{2,n} &: \mathcal{M}_{1,n} \times \mathcal{M}_{2,n} \to \hat{\mathcal{X}}^n. \end{aligned} \tag{29}$$

The first decoder receives only the base message and produces $\hat{\mathbf{X}}_1^n$. The second receives both messages and produces the finer reconstruction $\hat{\mathbf{X}}_2^n$. If R_1 and R_2 denote *cumulative* rates, then

$$\frac{1}{n} \log |\mathcal{M}_{1,n}| \leq R_1, \qquad \frac{1}{n} \log |\mathcal{M}_{2,n}| \leq R_2 - R_1. \tag{30}$$

Thus the second transmission contributes the incremental rate $R_2 - R_1$.

Definition 2. (Successive refinement from D_1 to D_2). *For $D_1 > D_2$, a source–distortion pair is successively refinable from D_1 to D_2 if the two-stage code can asymptotically attain*

$$R_1 = R(D_1), \qquad R_2 = R(D_2). \tag{31}$$

It is successively refinable in general if this is possible for every operational pair $D_1 > D_2$.

Condition Equation (31) is called the *no-excess-rate* condition. The refinement costs exactly $R(D_2) - R(D_1)$ additional bits per source symbol, which is the same total rate that would have been required had the fine distortion been known from the outset.

Single-letter rate region. The two-stage rate–distortion region can be characterized by a joint test channel $p(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2|\mathbf{x})$. A cumulative rate pair (R_1, R_2) is achievable if there exist reconstructions satisfying the distortion constraints and

$$R_1 \geq I(\mathbf{X}; \hat{\mathbf{X}}_1), \qquad R_2 \geq I(\mathbf{X}; \hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2). \tag{32}$$

Under no excess rate, both inequalities must align with the corresponding RD lower bounds.

Theorem 3. (Equitz–Cover characterization (Equitz & Cover, 1991)). *A source–distortion pair is successively refinable from D_1 to D_2 iff there exist reconstructions $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$ satisfying*

$$\begin{aligned} \mathbb{E}\left[d(\mathbf{X}, \hat{\mathbf{X}}_i)\right] &\leq D_i, \qquad i \in \{1, 2\}, \\ I(\mathbf{X}; \hat{\mathbf{X}}_i) &= R(D_i), \qquad i \in \{1, 2\}, \\ \mathbf{X} &\to \hat{\mathbf{X}}_2 \to \hat{\mathbf{X}}_1. \end{aligned}$$

The Markov condition is equivalent to the factorization

$$p(\mathbf{x}, \hat{\mathbf{x}}_2, \hat{\mathbf{x}}_1) = p(\mathbf{x})p(\hat{\mathbf{x}}_2|\mathbf{x})p(\hat{\mathbf{x}}_1|\hat{\mathbf{x}}_2). \tag{33}$$

It requires the coarse RD-optimal reconstruction to be a stochastic degradation of the fine RD-optimal reconstruction. It does not require $\hat{\mathbf{X}}_1$ to be a deterministic function of $\hat{\mathbf{X}}_2$.

Why no excess rate implies the Markov condition. Suppose a joint test channel attains the fine-stage no-excess rate,

$$I(\mathbf{X}; \hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2) = R(D_2), \tag{34}$$

while $\hat{\mathbf{X}}_2$ satisfies the fine distortion constraint. The RD converse and monotonicity of mutual information give

$$\begin{aligned} R(D_2) &\leq I(\mathbf{X}; \hat{\mathbf{X}}_2) \\ &\leq I(\mathbf{X}; \hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2) = R(D_2). \end{aligned} \tag{35}$$

All inequalities are therefore equalities. Applying the chain rule yields

$$I(\mathbf{X}; \hat{\mathbf{X}}_1|\hat{\mathbf{X}}_2) = 0, \tag{36}$$

which is precisely $\mathbf{X} \to \hat{\mathbf{X}}_2 \to \hat{\mathbf{X}}_1$.

B.5 Multistage successive refinement and nested codes

Multistage messages. For T stages, let $\mathbf{M}_k$ be the incremental message transmitted at stage k, with incremental rate r_k. Decoder k observes $\mathbf{M}_{1:k}$ and produces $\hat{\mathbf{X}}^{(k)}$. The cumulative rate is

$$R_k := \sum_{j=1}^{k} r_j, \qquad 0 < R_1 < \cdots < R_T. \tag{37}$$

The multistage achievable region admits a single-letter characterization in terms of a joint test channel $p(\hat{\mathbf{x}}^{(1)}, \ldots, \hat{\mathbf{x}}^{(T)}|\mathbf{x})$ (Rimoldi, 1994). In cumulative-rate notation, the constraints take the form

$$R_k \geq I\big(\mathbf{X}; \hat{\mathbf{X}}^{(1)}, \ldots, \hat{\mathbf{X}}^{(k)}\big), \qquad k = 1, \ldots, T, \tag{38}$$

together with the per-stage distortion constraints.

No excess rate at every prefix. A source–distortion pair is successively refinable at a finite allocation $\{R_k\}_{k=1}^{T}$ if every prefix attains its single-stage RD optimum:

$$D_k = D(R_k), \qquad k = 1, \dots, T. \tag{39}$$

Compatible optimal reconstructions can then be chosen to satisfy

$$\mathbf{X} \to \hat{\mathbf{X}}^{(T)} \to \hat{\mathbf{X}}^{(T-1)} \to \cdots \to \hat{\mathbf{X}}^{(1)}. \tag{40}$$

Under this chain, all earlier reconstructions provide no additional information about $\mathbf{X}$ once the finest currently available reconstruction is known. Consequently,

$$I\big(\mathbf{X}; \hat{\mathbf{X}}^{(1)}, \dots, \hat{\mathbf{X}}^{(k)}\big) = I(\mathbf{X}; \hat{\mathbf{X}}^{(k)}) = R(D_k). \tag{41}$$

Nestedness is not sufficient for SR. A nested code guarantees only that later decoders reuse earlier messages. It does not guarantee that every prefix is RD-optimal or that the reconstructions obey Equation (40). If the source–distortion pair is not SR at the chosen allocation, every nested code must satisfy

$$D_k > D(R_k) \quad \text{for at least one } k. \tag{42}$$

This is an incompatibility between individually optimal test channels, rather than a limitation that can be removed by increasing model capacity. A nested code may nevertheless be Pareto-optimal: no distortion can be reduced without increasing another distortion at the same cumulative rate allocation.

B.6 Gaussian sources under MSE

Gaussian RD function. Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m)$ and use the per-dimension MSE in Equation (18). For $0 < D \le \sigma^2$, the per-dimension RD and distortion–rate functions are (Cover & Thomas, 2006; Equitz & Cover, 1991)

$$R(D) = \frac{1}{2} \log_2 \frac{\sigma^2}{D}, \qquad D(R) = \sigma^2 2^{-2R}. \tag{43}$$

For $D \ge \sigma^2$, the zero-rate reproduction $\hat{\mathbf{X}} = \mathbf{0}$ is optimal.

An RD-optimal backward test channel can be written as

$$\mathbf{X} = \hat{\mathbf{X}} + \mathbf{N}, \qquad \hat{\mathbf{X}} \sim \mathcal{N}\big(\mathbf{0}, (\sigma^2 - D)\mathbf{I}_m\big), \qquad \mathbf{N} \sim \mathcal{N}(\mathbf{0}, D\mathbf{I}_m), \tag{44}$$

where $\hat{\mathbf{X}}$ and $\mathbf{N}$ are independent. This construction saturates the Shannon lower bound.

Equivalent additive forward channel. The same RD point can be realized from the noisy observation

$$\mathbf{Y} = \mathbf{X} + s\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m). \tag{45}$$

Joint Gaussianity gives the per-dimension information rate

$$\frac{1}{m} I(\mathbf{X}; \mathbf{Y}) = \frac{1}{2} \log_2 \left(1 + \frac{\sigma^2}{s^2}\right). \tag{46}$$

The MSE-optimal reconstruction is the conditional mean

$$\hat{\mathbf{X}} = \mathbb{E}[\mathbf{X}|\mathbf{Y}] = \frac{\sigma^2}{\sigma^2 + s^2} \mathbf{Y}, \tag{47}$$

with per-dimension distortion

$$D = \frac{\sigma^2 s^2}{\sigma^2 + s^2}. \tag{48}$$

Substitution into Equation (43) shows that $D = D(R)$. Conversely, a target rate R determines the required noise variance through

$$s^2 = \frac{\sigma^2}{2^{2R} - 1}. \tag{49}$$

Gaussian successive refinement. Choose cumulative rates $0 < R_1 < \cdots < R_T$ and corresponding noise variances from Equation (49). Then

$$s_1^2 > s_2^2 > \cdots > s_T^2. \tag{50}$$

The observations can be coupled as a degraded Gaussian chain. Starting from $\mathbf{Y}_T = \mathbf{X} + \mathbf{N}_T$, define

$$\mathbf{Y}_{k-1} = \mathbf{Y}_k + \mathbf{W}_k, \qquad \mathbf{W}_k \sim \mathcal{N}\big(\mathbf{0}, (s_{k-1}^2 - s_k^2)\mathbf{I}_m\big), \tag{51}$$

where all innovations are independent. This gives

$$\mathbf{X} \to \mathbf{Y}_T \to \cdots \to \mathbf{Y}_1. \tag{52}$$

At each level, $\hat{\mathbf{X}}^{(k)} = \sigma^2(\sigma^2 + s_k^2)^{-1}\mathbf{Y}_k$ is an invertible scaling of $\mathbf{Y}_k$. The Markov chain therefore transfers to the reconstructions, and every level attains $D(R_k)$. Gaussian sources under squared error are thus successively refinable (Equitz & Cover, 1991; Østergaard & Zamir, 2011).

The ordering in Equation (50) follows the manuscript's coarse-to-fine convention: level 1 has the largest noise, lowest rate, and highest distortion. This avoids reversing the refinement index and the forward noising-time index.

B.7 A NON-SR EXAMPLE: SYMMETRIC BINARY GAUSSIAN MIXTURES

Source and distortion regimes. Consider the scalar symmetric binary Gaussian mixture

$$\mathbf{X} \sim \frac{1}{2}\mathcal{N}(-\mu, \sigma^2) + \frac{1}{2}\mathcal{N}(+\mu, \sigma^2), \qquad \mu > 0, \tag{53}$$

and let the coordinates of the vector source be i.i.d. copies. The source is zero-mean with variance $\sigma^2 + \mu^2$, so its zero-rate distortion is

$$D_{\text{max}} = \sigma^2 + \mu^2. \tag{54}$$

For $0 < D \le \sigma^2$, the source admits the backward Gaussian decomposition required for equality in the Shannon lower bound. Hence

$$R(D) = h(\mathbf{X}) - \frac{1}{2}\log_2(2\pi e D), \qquad 0 < D \le \sigma^2. \tag{55}$$

The distortion $D = \sigma^2$ marks the critical boundary

$$R_{\text{crit}} := R(\sigma^2) = h(\mathbf{X}) - \frac{1}{2}\log_2(2\pi e \sigma^2). \tag{56}$$

In the intermediate range $\sigma^2 < D < \sigma^2 + \mu^2$, the optimal reproduction changes support and the RD function has no closed form used by this work. The reference frontier is therefore evaluated numerically using the mapping approach of Rose (1994).

Failure of successive refinement. Chow & Berger (1997) show that this source is not successively refinable whenever the coarse distortion lies in the intermediate regime and the fine distortion lies at or below the critical boundary:

$$\sigma^2 < D_1 < \sigma^2 + \mu^2, \qquad D_2 \le \sigma^2. \tag{57}$$

The RD-optimal test channels at these two distortions cannot be coupled so that $\mathbf{X} \to \hat{\mathbf{X}}_2 \to \hat{\mathbf{X}}_1$. Intuitively, the coarse representation and fine representation organize the mixture uncertainty differently; retaining an optimal coarse description as a prefix imposes a constraint that an independently designed fine description does not face.

The result in Equation (57) establishes a specific non-SR region. It does not, by itself, imply that every distortion pair outside this region is SR. Claims of refinability elsewhere require either an explicit Markov-compatible coupling of RD-optimal test channels or a separate theorem.

Table 3: Correspondence between classical lossy source coding and the hierarchical latent-variable models considered in this work.

Information-theoretic object	Model object	Role
Source symbol	$\mathbf{X}$	Data sample
Test channel	$q_\phi(\mathbf{Z}_{1:k}\|\mathbf{X})$	Stochastic encoder
Reproduction	$\hat{\mathbf{X}}^{(k)} = g_k(\mathbf{Z}_{1:k})$	Prefix reconstruction
Cumulative rate	$I(\mathbf{X};\mathbf{Z}_{1:k})$	Information in a prefix
Distortion	$\mathbb{E}[d(\mathbf{X},\hat{\mathbf{X}}^{(k)})]$	Reconstruction error
Refinement message	$\mathbf{Z}_k$	Increment beyond level $k-1$
SR compatibility	Reconstruction Markov chain	No excess prefix rate

B.8 Connection to hierarchical latent-variable models

Dictionary between source coding and latent models. The correspondence used throughout the paper is summarized in Table 3.

For a prefix reconstruction $\hat{\mathbf{X}}^{(k)} = g_k(\mathbf{Z}_{1:k})$, the model induces the Markov chain $\mathbf{X} \to \mathbf{Z}_{1:k} \to \hat{\mathbf{X}}^{(k)}$. Therefore,

$$D_k := \mathbb{E}\left[d(\mathbf{X},\hat{\mathbf{X}}^{(k)})\right] \geq D\big(I(\mathbf{X};\mathbf{Z}_{1:k})\big). \tag{58}$$

This bound holds independently at every level. SR asks the stronger question of whether equality can hold simultaneously for all prefixes of one nested code.

True and variational rates. For an HVAE prefix, define the aggregate posterior

$$q_\phi(\mathbf{z}_{1:k}) := \int p_{\text{data}}(\mathbf{x}) q_\phi(\mathbf{z}_{1:k}|\mathbf{x})\, d\mathbf{x}. \tag{59}$$

The expected KL used by variational training decomposes as

$$\begin{aligned} R_k^{\text{var}} &:= \mathbb{E}_{p_{\text{data}}(\mathbf{X})}\left[D_{\text{KL}}\big(q_\phi(\mathbf{Z}_{1:k}|\mathbf{X})\|p_\theta(\mathbf{Z}_{1:k})\big)\right] \\ &= \underbrace{I_{q_\phi}(\mathbf{X};\mathbf{Z}_{1:k})}_{R_k} + D_{\text{KL}}\big(q_\phi(\mathbf{Z}_{1:k})\|p_\theta(\mathbf{Z}_{1:k})\big) \\ &\geq R_k. \end{aligned} \tag{60}$$

The gap is the mismatch between the learned prior and the aggregate posterior. Consequently, $R_k^{\text{var}} = R_k$ when these distributions match.

Both rates admit hierarchical decompositions. The true information rate obeys the mutual-information chain rule

$$R_k = I(\mathbf{X};\mathbf{Z}_{1:k}) = \sum_{j=1}^{k} I(\mathbf{X};\mathbf{Z}_j|\mathbf{Z}_{1:j-1}), \tag{61}$$

whereas the factorization in Equation (2) gives

$$R_k^{\text{var}} = \sum_{j=1}^{k} \mathbb{E}_{p_{\text{data}}(\mathbf{X})q_\phi(\mathbf{Z}_{1:j-1}|\mathbf{X})}\Big[D_{\text{KL}}\big(q_\phi(\mathbf{Z}_j|\mathbf{X},\mathbf{Z}_{1:j-1})\|p_\theta(\mathbf{Z}_j|\mathbf{Z}_{1:j-1})\big)\Big]. \tag{62}$$

The j-th term is the variational incremental rate assigned to level j. Posterior collapse occurs when one or more such terms vanish because the decoder can reconstruct without those variables.

Since $D(R)$ is non-increasing and $R_k^{\text{var}} \geq R_k$, Equation (58) also implies

$$D_k \geq D(R_k) \geq D(R_k^{\text{var}}). \tag{63}$$

Thus comparison with $D(R_k^{\text{var}})$ remains a valid lower-bound diagnostic, but its gap can contain both reconstruction inefficiency and prior–aggregate-posterior mismatch. Keeping R_k and R_k^{var} notationally distinct makes this interpretation explicit.

Information rate versus realized code length. The mutual information $I(\mathbf{X};\mathbf{Z})$ is the asymptotic rate of a test channel under the source-coding theorem. It is not automatically the size of a stored floating-point latent. A literal compressor additionally requires quantization or a channel-simulation, entropy-coding, bits-back, or reverse-channel-coding construction. Similarly, the variational KL is a coding-motivated upper bound rather than a realized bitstream length unless an explicit coding scheme is supplied. This distinction does not affect the RD optimality analysis, which is expressed in information rates, but it matters when interpreting a model as a practical codec (Theis et al., 2022).

DDPM rate profiles. In the manuscript's coarse-to-fine notation, the prescribed DDPM encoder joint has the natural data-to-noise factorization

$$q(\mathbf{Z}_{1:T}|\mathbf{X}) = q(\mathbf{Z}_T|\mathbf{X}) \prod_{k=1}^{T-1} q(\mathbf{Z}_k|\mathbf{Z}_{k+1}). \tag{64}$$

The forward Markov chain gives

$$\begin{aligned} I(\mathbf{X};\mathbf{Z}_{1:k}) &= I(\mathbf{X};\mathbf{Z}_k) + I(\mathbf{X};\mathbf{Z}_{1:k-1}|\mathbf{Z}_k) \\ &= I(\mathbf{X};\mathbf{Z}_k). \end{aligned} \tag{65}$$

Hence a DDPM noise schedule directly fixes its cumulative information-rate profile. The numerical mapping from noise variance to rate depends on the source distribution: it is analytic for the Gaussian source but requires numerical evaluation for the Gaussian mixture.

Latent Markovity versus reconstruction Markovity. The Equitz–Cover condition concerns reconstructions, not latent states. From

$$\mathbf{X} \to \mathbf{Z}_k \to \mathbf{Z}_{k-1}, \tag{66}$$

and reconstructions $\hat{\mathbf{X}}^{(k)} = d_k(\mathbf{Z}_k;\theta)$, one cannot generally infer

$$\mathbf{X} \to \hat{\mathbf{X}}^{(k)} \to \hat{\mathbf{X}}^{(k-1)}. \tag{67}$$

A many-to-one denoiser may discard information in $\mathbf{Z}_k$ that remains relevant to the coarser state, so conditioning on $\hat{\mathbf{X}}^{(k)}$ need not screen off $\hat{\mathbf{X}}^{(k-1)}$ from the source. In the Gaussian/MSE case, the optimal denoiser is an invertible linear scaling, and the latent Markov chain does transfer to reconstructions. Outside that setting, reconstruction Markovity is an additional source-dependent condition rather than a generic consequence of the DDPM architecture.

C Proofs

C.1 Proof of Proposition 1

Assumptions and scope. Rates are measured in bits/dim and distortions in MSE/dim, following Section 2. We assume $\beta_k \in (0,1)$ for every diffusion level, interpret Equation (6) as a population $\mathbf{X}$-prediction loss with strictly positive weight at every level, and assume joint realizability of the Bayes denoisers: there exists a common $\theta^\star$ such that

$$d_k(\mathbf{Z}_k;\theta^\star) = \mathbb{E}[\mathbf{X}|\mathbf{Z}_k] \quad \text{almost surely}, \qquad k = 1,\dots,T. \tag{68}$$

The proposition concerns the autoencoding channel obtained by coupling the prescribed forward process with these denoisers, not equality between a learned reverse transition and the exact time reversal.

Proof. Because the paper indexes levels from coarse to fine, the forward process is the Markov chain

$$\mathbf{X} \to \mathbf{Z}_T \to \mathbf{Z}_{T-1} \to \cdots \to \mathbf{Z}_1. \tag{69}$$

Let $\bar{\alpha}_k := \prod_{j=k}^{T}(1-\beta_j)$. Recursive application of Equation (5) gives

$$\mathbf{Z}_k|\mathbf{X} \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_k}\mathbf{X}, (1-\bar{\alpha}_k)\mathbf{I}_m\right). \tag{70}$$

Define $\widetilde{\mathbf{Z}}_k := \mathbf{Z}_k/\sqrt{\bar{\alpha}_k}$ and $s_k^2 := (1-\bar{\alpha}_k)/\bar{\alpha}_k$. Then

$$\widetilde{\mathbf{Z}}_k|\mathbf{X} \sim \mathcal{N}\left(\mathbf{X}, s_k^2 \mathbf{I}_m\right). \tag{71}$$

The rescaling is invertible, and the forward Markov property gives $I(\mathbf{X}; \mathbf{Z}_{1:k}) = I(\mathbf{X}; \mathbf{Z}_k)$. Therefore the cumulative rate is

$$\begin{aligned} R_k &:= \frac{1}{m \log 2} I(\mathbf{X}; \mathbf{Z}_{1:k}) \\ &= \frac{1}{2} \log_2 \left(1 + \frac{\sigma^2}{s_k^2}\right). \end{aligned} \tag{72}$$

Since $\bar{\alpha}_{k-1} = (1-\beta_{k-1})\bar{\alpha}_k < \bar{\alpha}_k$, we have $s_{k-1}^2 > s_k^2$ and hence $0 < R_1 < \cdots < R_T$. Thus the noise schedule fixes an ordered cumulative-rate allocation independently of the denoiser.

For the Gaussian channel in Equation (71), the conditional mean is

$$\mathbb{E}[\mathbf{X}|\widetilde{\mathbf{Z}}_k] = a_k \widetilde{\mathbf{Z}}_k, \qquad a_k := \frac{\sigma^2}{\sigma^2 + s_k^2} \in (0,1). \tag{73}$$

Its per-dimension MSE is

$$\begin{aligned} D_k &= \frac{1}{m} \mathbb{E}\left[\|\mathbf{X} - a_k \widetilde{\mathbf{Z}}_k\|_2^2\right] \\ &= \frac{\sigma^2 s_k^2}{\sigma^2 + s_k^2} \\ &= \sigma^2 \left(1 + \frac{\sigma^2}{s_k^2}\right)^{-1} \\ &= \sigma^2 2^{-2R_k} = D(R_k), \end{aligned} \tag{74}$$

where the last equality is the Gaussian distortion–rate function. Hence the Bayes denoiser reaches the RD frontier at every schedule-induced rate.

It remains to show that every global minimizer recovers these denoisers. Writing $d_k^\star(\mathbf{Z}_k) := \mathbb{E}[\mathbf{X}|\mathbf{Z}_k]$, the orthogonality of conditional expectation gives

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{X} - d_k(\mathbf{Z}_k;\theta)\|_2^2\right] = \mathbb{E}\left[\|\mathbf{X} - d_k^\star(\mathbf{Z}_k)\|_2^2\right] \\ + \mathbb{E}\left[\|d_k^\star(\mathbf{Z}_k) - d_k(\mathbf{Z}_k;\theta)\|_2^2\right]. \end{aligned} \tag{75}$$

After summing over levels, joint realizability makes the first terms jointly attainable. Every global minimizer must therefore set every non-negative second term to zero, so $d_k(\mathbf{Z}_k;\theta) = d_k^\star(\mathbf{Z}_k)$ almost surely for all k.

Finally, Equation (73) and $\widetilde{\mathbf{Z}}_k = \mathbf{Z}_k/\sqrt{\bar{\alpha}_k}$ give $d_k^\star(\mathbf{Z}_k) = (a_k/\sqrt{\bar{\alpha}_k})\mathbf{Z}_k$, a nonzero linear rescaling. Each reconstruction consequently generates the same σ-algebra as its latent state. Replacing the variables in Equation (69) by these invertible transformations preserves conditional independence and yields

$$\mathbf{X} \to \hat{\mathbf{X}}^{(T)} \to \cdots \to \hat{\mathbf{X}}^{(1)}. \tag{76}$$

Together, Equations (74) and (76) give an RD-optimal multistage SR channel at the rates $\{R_k\}_{k=1}^T$. □

C.2 Proof of Proposition 2

Comparison class and assumptions. Let $\mathcal{C}$ be a class of model-realizable nested codes, with true rates $R_k(c)$ and distortions $D_k(c)$. We use the following idealized population assumptions: all expectations are exact; every prefix prior matches its aggregate posterior, so $R_k^{\text{var}}(c) = R_k(c)$; the prescribed schedule is valid; and the penalized population objective admits a global minimizer for every $\lambda > 0$. We additionally assume that the normalizers s_k are fixed independently of the optimized code and satisfy $0 < s_k < \infty$. The proof uses no other property of the reference curve from which they are obtained. Up to the constant factor $1/T$, the objective is therefore

$$J_\lambda(c) := \sum_{k=1}^T \frac{D_k(c)}{s_k} + \lambda \sum_{k=1}^T |R_k(c) - \gamma_k|. \tag{77}$$

Proof. Let c_λ be any global minimizer and write $r_k^{(\lambda)} := R_k(c_\lambda)$. Define the exactly feasible class and its optimal slope-normalized cumulative distortion by

$$\mathcal{F}_{\boldsymbol{\gamma}} := \{c \in \mathcal{C} : R_k(c) = \gamma_k \text{ for every } k\},$$
$$D^\star_{\boldsymbol{\gamma},\mathbf{s}} := \inf_{c \in \mathcal{F}_{\boldsymbol{\gamma}}} \sum_{k=1}^{T} \frac{D_k(c)}{s_k}. \tag{78}$$

Validity and positivity of the normalizers give $\mathcal{F}_{\boldsymbol{\gamma}} \neq \varnothing$ and $D^\star_{\boldsymbol{\gamma},\mathbf{s}} < \infty$. Every feasible code has zero rate penalty, so global optimality gives

$$\sum_{k=1}^{T} \frac{D_k(c_\lambda)}{s_k} + \lambda \sum_{k=1}^{T} |r_k^{(\lambda)} - \gamma_k| \leq D^\star_{\boldsymbol{\gamma},\mathbf{s}}. \tag{79}$$

Dropping the non-negative normalized distortion term yields

$$\sum_{k=1}^{T} |r_k^{(\lambda)} - \gamma_k| \leq \frac{D^\star_{\boldsymbol{\gamma},\mathbf{s}}}{\lambda} = O(\lambda^{-1}). \tag{80}$$

Now let $c' \in \mathcal{C}$ have the same achieved allocation as c_λ. The two codes incur the same rate penalty. Global optimality therefore implies

$$\sum_{k=1}^{T} \frac{D_k(c_\lambda)}{s_k} \leq \sum_{k=1}^{T} \frac{D_k(c')}{s_k}. \tag{81}$$

Thus, c_λ minimizes slope-normalized cumulative distortion at its achieved rates. If another code with those rates weakly improved every D_k and strictly improved at least one, positivity of every s_k would make its slope-normalized cumulative distortion strictly smaller, contradicting Equation (81). Hence c_λ is Pareto-optimal in the fixed-rate comparison class. □

C.3 Proof of Corollary 2.1

Proof. Let c_λ be an arbitrary global minimizer and write $r_k := R_k(c_\lambda)$. First, the achieved allocation is ordered for all sufficiently large λ. Indeed, let

$$\delta_{\boldsymbol{\gamma}} := \min\{\gamma_1, \gamma_2 - \gamma_1, \ldots, \gamma_T - \gamma_{T-1}\} > 0. \tag{82}$$

By Equation (80), the aggregate rate error is eventually smaller than $\delta_{\boldsymbol{\gamma}}$. Consequently, $r_1 > 0$ and, for $k \geq 2$,

$$r_k - r_{k-1} \geq \gamma_k - \gamma_{k-1} - |r_k - \gamma_k| - |r_{k-1} - \gamma_{k-1}| > 0. \tag{83}$$

Fix any λ for which $0 < r_1 < \cdots < r_T$.

Every nested code c with this allocation satisfies the per-level RD converse

$$D_k(c) \geq D(r_k), \qquad k = 1, \ldots, T. \tag{84}$$

Suppose first that the pair is SR at $\{r_k\}_{k=1}^T$ and that the model class contains a code $\bar{c}$ satisfying $R_k(\bar{c}) = r_k$ and $D_k(\bar{c}) = D(r_k)$ for every k. By the fixed-rate optimality in Proposition 2,

$$\sum_{k=1}^{T} \frac{D_k(c_\lambda)}{s_k} \leq \sum_{k=1}^{T} \frac{D(r_k)}{s_k}. \tag{85}$$

Combining this inequality with the RD converse gives

$$\sum_{k=1}^{T} \frac{D_k(c_\lambda) - D(r_k)}{s_k} = 0. \tag{86}$$

Every summand is non-negative because $s_k > 0$, so every summand must vanish. Hence $D_k(c_\lambda) = D(r_k)$ for every k. Since c_λ was arbitrary, the same conclusion holds for every global minimizer with this allocation.

Conversely, suppose the pair is not SR at $\{r_k\}_{k=1}^T$. If a nested code with these rates satisfied $D_k = D(r_k)$ for every k, it would provide a single nested code attaining the RD frontier at the full allocation, contradicting non-refinability. Hence every nested code with this allocation has $D_k > D(r_k)$ for at least one level. □

D ANALYTIC EXPERIMENTS

This appendix documents the analytic arm of Section 4.1 in full: the two sources and their reference curves (Appendix D.1), the model family and capacity accounting (Appendix D.2), the rate schedules and the DDPM calibration that makes the comparison rate-matched (Appendix D.3), the training protocol (Appendix D.4), every metric reported in the paper together with its bias sources (Appendix D.5), the complete rate–distortion and generative numbers (Appendices D.6 and D.7), and the λ ablation mentioned in Section 4.1 (Appendix D.8).

D.1 SOURCES AND REFERENCE CURVES

Both sources are 64-dimensional with i.i.d. coordinates, and distortion is per-dimension MSE as in Equation (18). All rates are in bits per source dimension.

Isotropic Gaussian. $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{64})$ with $\sigma = 1$. Its distortion–rate function is closed form, $D(R) = \sigma^2 2^{-2R}$, and the source is successively refinable under MSE at every allocation (Appendix B.6). Projected quantiles used by the SW2 estimators are exact, since $\mathbf{u}^\top \mathbf{X} \sim \mathcal{N}(0, \sigma^2)$ for any unit $\mathbf{u}$, and the MMD source-side expectations are available in closed form. The Gaussian source therefore introduces no source-side estimation error into any reported quantity.

Symmetric binary Gaussian mixture. $\mathbf{X}$ has i.i.d. coordinates drawn from $\frac{1}{2}\mathcal{N}(-\mu, \sigma^2) + \frac{1}{2}\mathcal{N}(+\mu, \sigma^2)$ with $\mu = 2$, $\sigma = 1$, so the per-coordinate variance is $\sigma^2 + \mu^2 = 5$ and $D_{\max} = 5$. Three distortion regimes are dispatched separately, following Appendix B.7:

- *subcritical*, $D \le \sigma^2$: the Shannon lower bound is tight and $R(D) = h(\mathbf{X}) - \frac{1}{2}\log_2(2\pi e D)$ in closed form;
- *supercritical*, $\sigma^2 < D < \sigma^2 + \mu^2$: no closed form, tabulated numerically;
- *degenerate*, $D \ge \sigma^2 + \mu^2$: $R(D) = 0$.

The boundary between the first two regimes defines the critical rate,

$$R_{\mathrm{crit}} := R(\sigma^2) = h(\mathbf{X}) - \tfrac{1}{2}\log_2\big(2\pi e \sigma^2\big) = 0.913 \text{ bits/dim}, \tag{87}$$

which is the rate at which Chow & Berger (1997) locate the failure of successive refinement: the pair is not refinable whenever the coarse distortion lies in the supercritical range and the fine distortion at or below σ^2. This is the quantity marked in Figure 1b.

The supercritical branch is precomputed once with the deterministic-annealing mapping algorithm of Rose (1994), sweeping 2000 inverse temperatures log-spaced over $\beta \in [10^{-3}, 15]$, with the source density resolved on a 2048-point quadrature grid spanning $\pm 8\sigma$, Lagrangian convergence tolerance 10^{-8}, centroid pruning below mass 10^{-5}, at most 10^4 centroid iterations per β, and bifurcation detection by symmetric splitting at perturbation 10^{-3} followed by 200 refinement iterations. The resulting table is stored once and read by every evaluation; $D(R)$ and $R(D)$ are obtained from it by shape-preserving PCHIP interpolation, which introduces the only source-side interpolation bias in the SBGM results.

D.2 MODELS AND CAPACITY

All arms use $T = 16$ levels. The two HVAE arms on each source are *exactly* capacity-matched: SR-HVAE differs from the vanilla HVAE only in the objective it minimises, not in a single parameter. Table 4 reports the architectures.

Both DDPMs predict a preconditioned $\mathbf{x}_0$ residual rather than $\boldsymbol{\epsilon}$, which is numerically better conditioned at the extreme noise levels the $[0.01, 3.0]$ rate range requires; the external $\boldsymbol{\epsilon}$ interface of Equation (6) is recovered exactly as $(\mathbf{z}_k - \hat{\mathbf{x}}_0)/s_k$. The SBGM denoisers additionally rescale their input by $\sqrt{\sigma_{\text{data}}^2 + s_k^2}$ before the nonlinearity.

On SBGM the HVAE family uses a coordinate-exchangeable backbone and a per-coordinate mixture prior with $K = 4$ equally weighted components. The mixture is required because a factorised

Table 4: Analytic-source architectures. HVAE and SR-HVAE are parameter-identical within each source; they differ only in whether principles (i)–(ii) are active. The Gaussian $T{=}1000$ denoiser is linear, which is exactly the Bayes-optimal family for that source under MSE and is why it needs so few parameters.

Source	Model	Parameters	Architecture
Gaussian	DDPM $T{=}1000$	321	linear denoiser, time embedding 64
	DDPM $T{=}16$	6,304	MLP, width 32, 1 main layer, time embedding 64
	HVAE $T{=}16$	136,528	MLP, $d_z{=}16$, encoder width 64 (1 layer),
	SR-HVAE $T{=}16$	136,528	decoder width 128 (2 layers), tiled readout, orthogonal input init
SBGM	DDPM $T{=}1000$	93,088	exchangeable MLP, width 96, 8 main layers
	DDPM $T{=}16$	93,088	same
	HVAE $T{=}16$	132,271	exchangeable backbone, $d_z{=}64$, 32 channels,
	SR-HVAE $T{=}16$	132,271	width 80, 3 layers, $K{=}4$ prior components

Gaussian prior cannot represent a bimodal aggregate posterior, which would confound a prior-expressiveness failure with the allocation failure the experiment is designed to isolate. Component spacing is solved once so that the prior attains a fixed initial KL floor, and the per-level posterior scale is initialised by bisection to overshoot the first incremental rate target by 5%.

D.3 Rate schedules and DDPM calibration

Every model targets the same cumulative schedule over $[\, r_{\min}, r_{\max}\,] = [0.01, 3.0]$ bits/dim,

$$\gamma_k = r_{\min} + \left(\tfrac{k-1}{T-1}\right)^{\alpha} (r_{\max} - r_{\min})\,, \qquad k = 1, \ldots, T, \tag{88}$$

with $\alpha = 1$ throughout the analytic experiments, so γ is linear in k. SR-HVAE receives γ directly through the penalty of Equation (10). The DDPMs receive it through their noise schedule, which must be inverted from rate to noise variance.

On the Gaussian source the inversion is analytic: $R_k = \frac{1}{2}\log_2(1+\sigma^2/s_k^2)$ gives $s_k^2 = \sigma^2/(2^{2\gamma_k}-1)$. On SBGM no closed form exists, and the mutual information $I(\mathbf{X}; \mathbf{X} + s\boldsymbol{\epsilon})$ is evaluated by Gauss–Legendre quadrature and inverted by bisection to a tolerance of 10^{-6} bits/dim per level. The resulting noise variances are stored once and loaded by every SBGM DDPM run.

This calibration is what makes the comparison in Section 4.1 rate-matched rather than merely similar: both families are asked for the same cumulative information at the same level index, and the measured rate gap of the DDPMs is 0.0000 on both sources because their rates are fixed by construction rather than optimised. The vanilla HVAE receives no schedule, since the ELBO of Equation (4) provides no mechanism to express one; its achieved rates are whatever training produces.

D.4 Training protocol

All analytic runs use AdamW at learning rate 10^{-4}, global batch size 1024, no gradient clipping, and a budget of at most 3×10^5 iterations with early stopping: the validation signal is evaluated every 2048 iterations on 10^4 fresh samples, and training halts after 4 consecutive evaluations without a relative improvement of at least 10^{-3}. Every configuration is run for seeds 0 through 9.

Every analytic run stopped early, so the iteration counts in Table 5 are convergence points rather than a fixed budget. SR-HVAE consistently takes longer to converge than the vanilla HVAE, which is expected: its objective constrains T cumulative rates and T prefix reconstructions rather than one of each. Because stopping is by convergence and not by a step count, this is a statement about optimisation difficulty and not a difference in training budget between the arms.

D.5 Metrics

All evaluations use a fixed evaluation seed of 4219, independent of the training seed, so that seed-to-seed variation in the reported numbers reflects training and not evaluation sampling.

Table 5: Iterations to early stopping across the 10 seeds of each analytic configuration.

Source	Model	min	max
Gaussian	DDPM T=1000	18,432	49,152
	DDPM T=16	10,240	22,528
	HVAE T=16	43,008	79,872
	SR-HVAE T=16, λ=1.0	49,152	94,208
	SR-HVAE T=16, λ=2.0	77,824	118,784
	SR-HVAE T=16, λ=4.0	94,208	131,072
SBGM	DDPM T=1000	10,240	18,432
	DDPM T=16	10,240	18,432
	HVAE T=16	22,528	57,344
	SR-HVAE T=16, λ=1.0	32,768	71,680

Rate targeting (principle (i)). $\frac{1}{T}\sum_k |\gamma_k - R_k|$, with R_k family-specific: the cumulative variational KL of Equation (7) for the HVAE family, and the forward-channel mutual information for the DDPMs, which is closed form on the Gaussian source and read from the calibration table on SBGM. Evaluated on 25,000 samples.

Distortion gap (principle (ii)). $D_k - D(R_k)$, evaluated at the model's *own achieved* rate rather than at γ_k, so that a failure to hit the rate target does not contaminate the distortion diagnostic. We report it in normalised form as the excess ratio $D_k/D(R_k) \geq 1$. Source-side error is nil on the Gaussian source and limited to PCHIP interpolation on SBGM.

Sliced Wasserstein-2. $\mathrm{SW}_2^2(p,q) = \mathbb{E}_{\mathbf{u}}[W_2^2(P_{\mathbf{u}}\#p, P_{\mathbf{u}}\#q)]$ with the one-dimensional W_2 evaluated on a common quantile grid, the model side from sorted samples and the source side analytically. Two projection sets are used, and the distinction matters:

- **rp-SW2** averages over 1024 random unit directions. Because the 64 coordinates are i.i.d., $\mathbf{u}^\top\mathbf{X}$ concentrates towards a Gaussian by the central limit theorem, with projected excess kurtosis $\kappa_4(\mathbf{u}^\top\mathbf{X}) \approx \kappa_4(X)/m$ and $\kappa_4(X) = -2\mu^4/(\sigma^2+\mu^2)^2$. On SBGM this washes out the per-coordinate bimodality entirely, so rp-SW2 is *not* informative about whether a model reproduced the mixture; it remains a valid diagnostic for cross-dimensional coupling.
- **ax-SW2** averages over the 64 coordinate axes, so it compares per-coordinate marginals directly and is sensitive to mode count and mode geometry. It is the primary SBGM distributional metric. On the Gaussian source the two coincide to within sampling noise, which is why the main text reports ax-SW2 only.

MMD. Squared maximum mean discrepancy under a Gaussian kernel, reported at bandwidth h_{med}, the median heuristic computed once from 50,000 source samples and frozen across every evaluation in the paper. The five multipliers $\{0.25, 0.5, 1, 2, 4\}\, h_{\text{med}}$ are computed from the same pairwise-distance matrix at negligible extra cost and are reported in the released result files. Source-side kernel expectations are closed form.

Prior-slack bracket. The quantity $\delta_k = R_k^{\text{var}} - R_k$ of Equation (8) is not directly computable, because R_k requires the aggregate posterior density. We bracket it by training auxiliary normalised density models on fresh trajectories and evaluating them on independent batches: a variational upper bound and a Barber–Agakov lower bound on R_k, each taken as the best of a ladder of three density classes (diagonal Gaussian, mixture of 8 Gaussians, and an autoregressive mixture). Each rung is trained for 10^4 steps at batch 8192, learning rate 5×10^{-3}, width 128, 2 hidden layers, and evaluated on 65,536 samples. The reported bracket $[\,\delta_T^{\text{lo}}, \delta_T^{\text{hi}}\,]$ is therefore an estimate whose width reflects the auxiliary estimator's own slack; a narrower bracket is a stronger statement. Both SR-HVAE arms reported in Table 1 use identical estimator settings, so their brackets are directly comparable. The quantity is defined only for a learned encoder: for a prescribed one, R_k is a functional of the source and the noise schedule alone and needs no variational bound.

Table 6: Rate–distortion results on the analytic sources, mean (standard deviation) over 10 seeds. R_T is the cumulative rate reached at the finest level, against a target of 3.0 bits/dim; the rate gap is $\frac{1}{T}\sum_k |\gamma_k - R_k|$; the excess ratio is $D_k/D(R_k)$ at the achieved rate, reported as the median and the maximum over levels. The DDPM rate gap is zero by construction (Appendix D.3).

Source	Model	R_T	rate gap	median $D_k/D(R_k)$	$\max_k D_k/D(R_k)$
Gaussian	DDPM T=16	3.000 (0.000)	0.0000 (0.0000)	1.000 (0.000)	1.002 (0.000)
	HVAE T=16	0.207 (0.004)	1.3844 (0.0030)	1.052 (0.003)	1.061 (0.006)
	SR-HVAE T=16	3.006 (0.001)	0.0391 (0.0013)	1.080 (0.004)	1.102 (0.004)
SBGM	DDPM T=16	3.000 (0.000)	0.0000 (0.0000)	1.101 (0.000)	1.330 (0.000)
	HVAE T=16	0.577 (0.608)	1.1541 (0.3718)	1.068 (0.078)	1.152 (0.178)
	SR-HVAE T=16	3.035 (0.016)	0.0291 (0.0091)	1.005 (0.002)	1.080 (0.007)

Table 7: Per-seed outcome of the vanilla HVAE on SBGM. $D_{\max} = 5$ is the zero-rate distortion.

seed	0	1	2	3	4	5	6	7	8	9
R_T	$8{\cdot}10^{-6}$	1.151	1.152	1.153	$4{\cdot}10^{-5}$	1.160	1.156	$7{\cdot}10^{-6}$	$3{\cdot}10^{-6}$	$2{\cdot}10^{-5}$
D_T	5.002	0.722	0.721	0.722	5.002	0.713	0.717	5.002	5.002	5.002
ax-SW2	2.236	0.439	0.437	0.437	2.236	0.442	0.438	2.236	2.236	2.236

D.6 Complete rate–distortion results

Table 6 gives the numbers underlying Figure 1.

Reading the vanilla-HVAE rows. The excess ratios of the vanilla HVAE are evaluated over a far smaller rate range than the other rows, since it never approaches the target, and are therefore not comparable to them: a model that stops at $R_T = 0.21$ is being measured only on its first few levels. Its standard deviations on SBGM are also not noise but bimodality. Table 7 lists the per-seed outcome: 5 of 10 seeds collapse to $R_T \approx 10^{-5}$ with $D_k = D_{\max}$ at *every* level, and the remaining 5 stall at $R_T \approx 1.15$. A mean and standard deviation over that population describes neither mode, which is why Section 4.1 reports the collapse frequency instead, and why the collapsed seeds are excluded from the HVAE curve in Figure 1b. Counting prefixes that carry no usable information makes the same point without averaging: on SBGM, 8 of the vanilla HVAE's 16 prefixes reconstruct no better than the source mean, against 0 of 16 for SR-HVAE, the T=16 DDPM, and every arm on the Gaussian source.

D.7 Complete generative results and the per-level prior slack

Table 8 extends Table 1 with rp-SW2 and with the standard deviations omitted there for space.

Table 8: Generative metrics on the analytic sources, mean (standard deviation) over 10 seeds. rp-SW2 is uninformative about per-coordinate bimodality on SBGM, for the reason given in Appendix D.5; ax-SW2 is the primary metric there.

Source	Model	rp-SW2	ax-SW2	MMD
Gaussian	DDPM T=1000	0.0154 (0.0003)	0.0157 (0.0003)	$3.2{\cdot}10^{-5}$
	DDPM T=16	0.1348 (0.0005)	0.1349 (0.0005)	$6.6{\cdot}10^{-3}$
	HVAE T=16	0.5070 (0.0049)	0.5069 (0.0049)	$7.9{\cdot}10^{-2}$
	SR-HVAE T=16	0.0210 (0.0009)	0.0210 (0.0009)	$1.1{\cdot}10^{-4}$
SBGM	DDPM T=1000	0.0333 (0.0004)	0.5144 (0.0000)	$4.3{\cdot}10^{-5}$
	DDPM T=16	0.3096 (0.0001)	0.5698 (0.0000)	$7.0{\cdot}10^{-3}$
	HVAE T=16	1.2049 (1.0868)	1.3372 (0.9474)	$8.0{\cdot}10^{-2}$
	SR-HVAE T=16	0.0352 (0.0042)	0.0372 (0.0039)	$2.9{\cdot}10^{-5}$

Table 9: Bracket on the prior slack δ_k (bits/dim) at every level, mean over 10 seeds, for SR-HVAE at $\lambda = 1$. Upper row of each block is δ_k^{hi}, lower is δ_k^{lo}.

Source		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Gaussian	hi	.0016	.0032	.0082	.0101	.0129	.0157	.0150	.0138	.0163	.0217	.0256	.0253	.0271	.0310	.0324	.0386
	lo	.0004	.0007	.0011	.0013	.0014	.0010	.0013	.0012	.0012	.0018	.0025	.0031	.0041	.0056	.0071	.0089
SBGM	hi	.0006	.0080	.0119	.0146	.0159	.0151	.0057	.0042	.0035	.0031	.0030	.0032	.0036	.0041	.0047	.0055
	lo	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Table 10: Effect of λ on the gaussian source, mean over 10 seeds. Increasing λ tightens rate targeting and loosens the per-prefix distortion, exactly the exchange Equation (10) encodes.

Source	λ	R_T	rate gap	$\max_k \lvert\gamma_k - R_k\rvert$	$\max_k D_k/D(R_k)$	ax-SW2
Gaussian	1.0	3.006	0.0391	0.0825	1.102	0.0210
	1.5	3.003	0.0079	0.0608	1.113	0.0210
	2.0	3.003	0.0065	0.0488	1.120	0.0214
	4.0	3.001	0.0022	0.0099	1.160	0.0235

The SBGM block makes the metric distinction concrete. Under random projections the $T{=}1000$ DDPM appears excellent (0.0333) and essentially matches SR-HVAE (0.0352); under axis-aligned projections it is 0.5144 against SR-HVAE's 0.0372, a factor of 14. The DDPM values are identical to four decimal places across all 10 seeds, so this is a systematic property of the prescribed Gaussian channel on a mixture source and not seed variation. The two metrics agree on the Gaussian source for every arm, which rules out an artifact of the estimator.

Per-level prior slack. Table 9 gives the bracket on δ_k at every level, in bits/dim. On the Gaussian source the slack accumulates monotonically with depth, reaching at most 0.0386 bits/dim at $k = 16$, which is 1.28% of R_T^{var}. On SBGM it peaks in the middle of the hierarchy and then falls, ending at 0.0055 bits/dim, or 0.18%. In both cases the variational rate the objective controls is within a couple of percent of the true rate, which is the empirical content of the variational-tightness assumption of Proposition 2.

D.8 The rate–distortion exchange λ

The main text keeps $\lambda = 1$, which by the construction of Equation (10) matches the local reference trade-off: one bit/dim of rate deviation costs exactly as much as the distortion reduction the frontier would grant for it. Larger λ prioritises rate enforcement. Table 10 sweeps it.

The trend is monotone. Raising λ from 1 to 4 on the Gaussian source reduces the mean rate error by a factor of 18 and the worst-level rate error by a factor of 8, while the worst-level distortion excess rises from 1.102 to 1.160 and ax-SW2 degrades from 0.0210 to 0.0235. This is the behaviour Proposition 2 describes: the rate penalty is a relaxation whose aggregate error vanishes as $\lambda \to \infty$, and the price is paid in the slope-normalised distortion the second term minimises at the achieved allocation. We use $\lambda = 1$ in the main text because the balanced setting is the one with a source-independent interpretation, and because the generative metrics are best there gaussian source.

E Image experiments

This appendix documents the image experiments of Section 4.2 in full: the two datasets and the reference curves they are scored against (Appendix E.1), the architectures and capacity accounting (Appendix E.2), the rate schedules and the DDPM calibration that makes the comparison rate-matched (Appendix E.3), the training protocol together with the evidence that the ELBO's rate choice is structural (Appendix E.4), every metric and sampling arm (Appendix E.5), the complete rate–distortion and generative numbers including the α ablation mentioned in Section 4.2 (Appendices E.6 and E.7), the prior-slack measurements and the post-hoc refit (Appendix E.8), and the prefix diagnostics and held-out results (Appendix E.9).

Table 11: Image architectures. Within each dataset the vanilla HVAE and the SR-HVAE are parameter-identical and differ only in whether principles (i)–(ii) are active; the DDPM is a separate backbone and is *smaller* than both.

Dataset	Model	Parameters	Architecture
MNIST	HVAE T=32	1,940,677	conv encoder, d_z=32, latent 1×8×8,
	SR-HVAE T=32	1,940,677	base 40 ch., mult. $(1, 2, 4)$, shared FiLM
	DDPM T=32/1000	1,300,447	U-Net, base 17 ch., 2 res. blocks, attn. 8
CelebA	HVAE T=32	5,004,227	conv encoder, d_z=512, latent 2×16×16,
	SR-HVAE T=32	5,004,227	base 56 ch., mult. $(1, 2, 4)$, shared FiLM
	DDPM T=32/1000	2,730,691	U-Net, base 27 ch., 2 res. blocks, attn. 16

E.1 Datasets and reference curves

MNIST digits are resized to 32×32 and kept single-channel, giving $m = 1024$ source dimensions. CelebA is centre-cropped to the 128×128 window at $(89, 121)$ used by the public DDIM checkpoint and resized to 64×64, giving $m = 12288$; the training split additionally receives random horizontal flips with probability 0.5, and the test split receives none. Pixels are mapped to $[-1, 1]$ and distortion is per-dimension MSE in that space, so the zero-rate distortion is $D_{\max} = \sigma^2_{\text{data}}$, equal to 0.3348 on MNIST and 0.2893 on CelebA. A prefix whose distortion reaches $D_{\max}$ carries no usable information about its input: the data mean would reconstruct as well.

Unlike the analytic sources, images have no closed-form $D(R)$. We therefore score against an *achievable* curve obtained from the I-MMSE relation (Guo et al., 2005), which states that the derivative of the mutual information of a Gaussian channel with respect to its signal-to-noise ratio is half the minimum mean-square error at that ratio. Integrating a reference denoiser's MMSE over the noise range therefore yields a (R, D) table that some encoder–decoder pair attains, and which upper bounds the unknown $D(R)$. Two consequences matter for reading our tables. First, a model may legitimately fall *below* the reference, and the SR-HVAE does so at a few levels on MNIST; the code retains the signed gap on image sources for this reason, where it clamps it on analytic ones. Second, the reference is generated by a Gaussian channel, which is exactly the encoder family a DDPM uses, so a DDPM sitting near the reference is close to tautological and should not be read as evidence of optimality. The reference denoiser is independent of every compared arm: on MNIST it is a T=1000 DDPM trained for $200{,}000$ iterations for this purpose alone, on CelebA a public DDIM 64×64 checkpoint. Quadrature is validated against the analytic Gaussian bound to 10^{-4} bits and against monotonicity of the MMSE sequence under a one-sided Bonferroni correction at family-wise level 0.05.

E.2 Models and capacity

Table 11 lists the three families. The comparison that carries the paper's first claim — vanilla HVAE against SR-HVAE — is exactly capacity-matched, since the two share a configuration file up to the `principles` and `alpha` fields. The DDPM is a different backbone and cannot be matched parameter-for-parameter; but the order of magnitude is the same as the genrative path of HVAEs. Both DDPMs predict a preconditioned $\mathbf{x}_0$ residual with signal-to-noise input scaling, as on the analytic sources. The post-hoc refit of Appendix E.8 replaces the conditional prior with a U-Net of base width 40 (MNIST) or 56 (CelebA) and context-dependent log-variance, which is the only place any architecture changes.

E.3 Rate schedules and DDPM calibration

All arms use the canonical power schedule $\gamma_k = r_{\min} + (k/T)^\alpha \, (r_{\max} - r_{\min})$ with $T = 32$. The floor $r_{\min}$ is the first incremental rate the vanilla HVAE produces on its own, 9.730×10^{-5} on CelebA and 4.129×10^{-4} on MNIST, so that no arm is asked to open at a rate the family cannot reach. The ceiling is the experimental variable: the matched budget sets $r_{\max}$ to the vanilla HVAE's own achieved total (0.03246 and 0.01317 bits/dim), and the extended budget raises it to 0.2 and 0.1 bits/dim. The exponent α redistributes the same budget: $\alpha < 1$ front-loads it, so most levels operate

in the high-rate region, while $\alpha > 1$ back-loads it and reserves most levels for the low-rate region where $|D'_{\text{ref}}(R)|$ is largest. At $\alpha = 2$ the schedule has committed 5.2% of the budget by level 8; at $\alpha = 1$, 22.7%; at $\alpha = 0.5$, 47.6%.

DDPM arms are rate-matched by construction rather than by optimization. The noise schedule is obtained by inverting the same I-MMSE table at each γ_k, so the cumulative rate a DDPM carries at level k equals the target to within the table's resolution; the measured rate-targeting gap is 0.00000 bits/dim for every DDPM arm in Table 12. This is what makes the schedule a knob shared by both families, and it is why the α ablation can be run on a model whose encoder we never train.

E.4 Training protocol

All models are trained with AdamW without gradient clipping, at learning rate 10^{-4} and batch size 256 on MNIST, and 7.07×10^{-5} and batch 128 on CelebA. The training iterations are $21{,}500$ for the matched-budget arms and all DDPMs, $43{,}000$ for the extended-budget SR-HVAEs on MNIST, and $101{,}700$ and $127{,}160$ for the corresponding CelebA arms (80 and 100 epochs). Longer training on SR-HVAEs is needed as they need to optimize an harder objective (two terms at T levels, agains DDPM's one terms and HVAE 1 level).

E.5 Metrics and sampling arms

FID and KID are computed with `torch-fidelity` against cached Inception-V3 `pool3` statistics over the complete training split, from $50{,}000$ generated samples; the `torch-fidelity` version and the reference-cache hash are recorded in every result file. Improved precision and recall (Kynkäänniemi et al., 2019) and density and coverage (Naeem et al., 2020) use the same features with a $k = 5$ nearest-neighbour manifold. Recall is known to be unstable when the generated manifold barely intersects the reference one, which is the regime every CelebA arm occupies (values between 0.004 and 0.198); we therefore report density and coverage alongside it in Table 13 and base no claim on recall alone.

Four sampling arms recur. The *ancestral* arm draws every latent from the generative prior and decodes the first k; it is the generative trajectory. The *reconstruction* arm drives every latent from the encoder, so it measures what the decoder can produce from the representation the model actually learned, and it upper bounds what any prior could deliver at the same rate and distortion. For a DDPM this arm is the forward-noising-and-denoising path, which is why it is defined for both families. The *hybrid-j* arm conditions the first j levels on the encoder and samples the remainder from the prior, decoding at level T; sweeping j localizes where along the hierarchy the prior slack sits. The *temperature* arm scales the prior standard deviation below or above one. All arms share the same seeded input indices within a dataset and split, which the result files record as a SHA-256 of the index tensor, so every number in this appendix is computed on the same images.

E.6 Complete rate–distortion results

Table 12 contains the numbers behind the first half of Section 4.2. Three readings are worth isolating. The vanilla HVAE is the only arm with failed prefixes, and its excess peaks in the *middle* of the hierarchy (level 16 on both datasets) rather than at either end, which is the signature of a model that is free to defer all useful reconstruction to the last levels. The SR-HVAE removes those failures at the matched rate with a median excess of 1.01 and 1.12, at identical capacity. Finally, the DDPM's excess is flat in α — between 1.11 and 1.27 across the whole sweep — even though its FID varies by a factor of 8.5 over the same sweep, so the schedule's effect on a prescribed encoder is not a frontier effect but a question of where its 32 steps are placed.

The vanilla HVAE's rate allocation *within* its region is not pathological: its per-level increments span 1.3% to 3.4% of the total on MNIST and 0.7% to 3.9% on CelebA, with no level below 0.5% and 29 of 32 levels needed to reach 90% of R_T. This distinguishes the failure from posterior collapse, where individual levels carry no rate at all, and from the collapse we do observe on the analytic SBGM source (Table 7). On images the increments are present and the reconstructions are not.

Table 12: Rate–distortion results at $T = 32$. The rate-targeting gap is the mean absolute deviation of R_k^{var} from γ_k; it is undefined for the vanilla HVAE, which is given no schedule. Excess is $D_k/D_{\text{ref}}(R_k)$, reported as the median and maximum over the 32 prefixes. The last column counts prefixes whose distortion reaches $D_{\max}$, *i.e.* that reconstruct no better than the data mean.

Dataset	Model	R_T	rate gap	excess (med.)	excess (max.)	$D_k \geq D_{\max}$
MNIST	HVAE	0.0324	—	2.14	3.37	$4/32$
	SR-HVAE (matched)	0.0329	0.00056	1.01	1.25	$0/32$
	DDPM (matched)	0.0325	0.00000	1.07	1.15	$0/32$
	SR-HVAE α=0.5	0.2013	0.02430	2.81	4.73	$0/32$
	SR-HVAE α=1	0.1995	0.01688	2.39	3.30	$0/32$
	SR-HVAE α=2	0.2000	0.01025	1.98	3.31	$0/32$
	DDPM α=0.5	0.2000	0.00000	1.11	1.23	$0/32$
	DDPM α=1	0.2000	0.00000	1.14	1.20	$0/32$
	DDPM α=2	0.2000	0.00000	1.16	1.27	$0/32$
CelebA	HVAE	0.0132	—	4.81	6.95	$12/32$
	SR-HVAE (matched)	0.0141	0.00052	1.12	1.29	$0/32$
	DDPM (matched)	0.0132	0.00000	1.10	1.18	$0/32$
	SR-HVAE α=0.5	0.1017	0.00727	2.09	4.24	$0/32$
	SR-HVAE α=1	0.1018	0.00500	1.80	2.29	$0/32$
	SR-HVAE α=2	0.1014	0.00323	1.58	2.28	$0/32$
	DDPM α=0.5	0.1000	0.00000	1.15	1.20	$0/32$
	DDPM α=1	0.1000	0.00000	1.17	1.18	$0/32$
	DDPM α=2	0.1000	0.00000	1.21	1.23	$0/32$

Table 13: Generative metrics at the final level, $50{,}000$ samples against the training split. The 1000-step DDPM is evaluated only at its final level. Density and coverage are reported for the arms that have prefix diagnostics; recall should be read together with them on CelebA, where all values are small.

Dataset	Model	FID ↓	KID ↓	prec. ↑	rec. ↑	cov. ↑
MNIST	HVAE	21.59	0.0196	0.237	0.288	0.230
	SR-HVAE (matched)	22.80	0.0212	0.294	0.236	0.265
	DDPM T=32 (matched)	46.75	0.0609	0.039	0.114	—
	DDPM T=1000 (matched)	43.06	0.0558	0.037	0.123	—
	SR-HVAE α=0.5	26.60	0.0241	0.314	0.640	0.210
	SR-HVAE α=1	13.57	0.0084	0.356	0.648	0.309
	SR-HVAE α=2	11.09	0.0067	0.379	0.632	0.364
	DDPM T=32, α=0.5	89.36	0.0993	0.103	0.131	—
	DDPM T=32, α=1	13.95	0.0152	0.502	0.628	—
	DDPM T=32, α=2	10.47	0.0106	0.539	0.662	—
	DDPM T=1000, $\alpha\in\{0.5, 1, 2\}$	8.38 / 5.61 / 7.84	—	—	—	—
CelebA	HVAE	48.26	0.0465	0.286	0.005	0.113
	SR-HVAE (matched)	53.14	0.0505	0.309	0.004	0.114
	DDPM T=32 (matched)	38.37	0.0343	0.436	0.028	—
	DDPM T=1000 (matched)	30.29	0.0236	0.522	0.037	—
	SR-HVAE α=0.5	105.54	0.1382	0.157	0.024	0.028
	SR-HVAE α=1	54.21	0.0645	0.397	0.013	0.123
	SR-HVAE α=2	50.79	0.0593	0.367	0.024	0.134
	DDPM T=32, α=0.5	60.40	0.0601	0.509	0.004	—
	DDPM T=32, α=1	38.09	0.0357	0.640	0.058	—
	DDPM T=32, α=2	30.84	0.0310	0.577	0.118	—
	DDPM T=1000, $\alpha\in\{0.5, 1, 2\}$	32.67 / 25.64 / 26.01	—	—	—	—

E.7 Complete generative results and the α ablation

The α ablation. Back-loading the schedule improves every T=32 arm on both datasets, in both families: the SR-HVAE moves from 26.60 to 11.09 on MNIST and 105.54 to 50.79 on CelebA as α goes from 0.5 to 2, and the DDPM from 89.36 to 10.47 and 60.40 to 30.84. That is four monotone

Table 14: Prior slack and its post-hoc removal. The sliced-KL column is $\sum_k D_{\mathrm{KL}}(q_\phi(\mathbf{Z}_k)\|p_\theta(\mathbf{Z}_k))$ estimated along random one-dimensional projections, as a percentage of R_T, and is a lower bound on δ_T. Variant A retrains the existing prior; variant B fits a larger U-Net prior. Both freeze the encoder and the decoder, so $\{R_k, D_k\}$ are unchanged and only the sampling path moves. FID is the ancestral value after the refit; only variant B was evaluated generatively.

			variational rate (val.)		ancestral FID	
Dataset	Model	sliced KL	trained	after refit	before	after (B)
MNIST	HVAE	0.97%	0.0324	0.0314	21.59	23.72
	SR-HVAE (matched)	0.96%	0.0330	0.0324	22.80	25.62
	SR-HVAE α=0.5	0.75%	0.2025	0.1973	26.60	15.01
	SR-HVAE α=2	0.27%	0.2001	0.1925	11.09	**6.46**
CelebA	HVAE	0.74%	0.0132	0.0129	48.26	48.22
	SR-HVAE (matched)	0.70%	0.0142	0.0140	53.14	51.53
	SR-HVAE α=0.5	0.41%	0.1018	0.0989	105.54	77.33
	SR-HVAE α=2	0.13%	0.1015	0.0990	50.79	**43.32**

sweeps out of four, and $\alpha = 2$ is the largest value we tested, so the ordering is monotone across the tested range rather than optimal at its endpoint.

Two observations locate the mechanism. First, it is not a frontier effect for the prescribed encoder: the DDPM's rate–distortion excess is flat across the sweep (Table 12) while its FID moves by $8.5\times$. Second, the effect is specific to a scarce step budget. At T=1000 the same sweep spans only 8.38–7.84 on MNIST and 32.67–26.01 on CelebA, a factor of 1.5 instead of 8.5, and the optimum moves back to $\alpha = 1$. With enough steps everywhere the placement stops mattering; with 32 it decides the result. This is what the objective's $1/s_k$ normalization encodes, and we stress that s_k and α act on different things: s_k equalizes the cost of a given rate error across levels, while α decides where the levels are placed in the first place.

E.8 Prior slack and the post-hoc refit

The refit is an intervention on δ alone. Every parameter whose name does not begin with the conditional-prior heads is frozen, which the implementation asserts, so the encoder, the decoder and therefore the achieved $\{R_k, D_k\}$ cannot move; what changes is only the distribution the generative path samples from. Variant A warm-starts the trained prior and retrains it against the cached posterior KL; variant B builds a fresh, larger U-Net prior and copies the frozen weights across. Both run for 20,000 steps on 50,000 cached posteriors with 5,000 held out, at learning rate 10^{-4} (A) and 3×10^{-4} (B).

Table 14 shows a small rate reduction with a large generative effect where the allocation was repaired first, and no effect at all where it was not. On the MNIST α=2 arm the validation rate falls by 0.0076 bits/dim, 3.8% of the total, and the ancestral FID falls by 42%; on the vanilla HVAE a comparable relative reduction changes nothing ($21.59 \rightarrow 23.72$ on MNIST, $48.26 \rightarrow 48.22$ on CelebA). The matched-rate SR-HVAE behaves like the vanilla HVAE rather than like its extended-budget sibling, so the correct reading is not that our objective makes priors refittable, but that a rate allocation worth carrying is a precondition for the refit to convert into samples.

Two measurement caveats. On MNIST variants A and B converge to the same validation rate (0.1925 both), which indicates a floor; on CelebA variant B reaches 0.0990 against A's 0.1007, so no floor was reached and the CelebA entries are loose lower bounds on δ_T. Separately, only variant B was evaluated generatively, so the reported FID improvements are attributable to the refit *and* the larger prior together; and the refit prior is fitted to cached training posteriors and scored against training-split statistics, while the base model's train-to-test offset is $+0.73$ FID (Table 15).

Table 15: Prefix diagnostics for the SR-HVAE. Hybrid-j conditions the first j of 32 levels on the encoder and samples the rest from the prior, decoding at level T; it interpolates between the ancestral and reconstruction arms and localizes where the prior slack sits. The held-out columns repeat the two end arms on the test split. Dashes mark arms absent from the evaluation configuration used for that run.

		train		hybrid-j (train)						held out	
Dataset	Model	anc.	rec.	1	2	4	8	16	24	anc.	rec.
MNIST	matched	22.80	26.70	22.76	22.85	23.09	23.55	24.81	25.89	23.40	27.83
	α=0.5	26.60	2.95	14.17	5.82	3.62	3.27	3.09	3.01	27.36	4.02
	α=1	13.57	3.35	12.54	10.24	6.43	4.35	4.08	3.71	14.35	4.41
	α=2	11.09	3.36	10.96	10.55	9.83	8.11	4.78	4.57	11.82	4.42
CelebA	matched	53.16	41.28	52.26	51.23	50.04	47.99	45.11	42.82	53.53	42.80
	α=0.5	105.78	26.25	62.84	44.86	35.54	29.66	27.25	26.57	105.74	27.44
	α=1	54.21	23.73	51.56	—	44.65	35.56	26.97	—	54.49	24.86
	α=2	50.66	24.21	50.48	—	49.72	46.71	35.31	—	50.98	25.34

E.9 Prefix diagnostics and held-out results

The hybrid-j columns of Table 15 read the schedule back out of the trained model. Under front-loading, conditioning on a single level already halves the MNIST FID ($26.60 \to 14.17$) and two levels bring it to 5.82: almost all of the prior slack sits at the first level, which is exactly where the schedule placed the rate. Under back-loading the same curve is gradual (10.96 at j=1, 8.11 at j=8, 4.78 at j=16), because no single level carries enough rate for its prior to dominate. The matched-rate arm is the instructive exception: its hybrid curve *rises* with j, since at 0.033 bits/dim the reconstruction arm (26.70) is worse than the ancestral one (22.80), and conditioning on the encoder moves samples towards a representation that is itself the limitation.

The held-out columns confirm that none of this is memorisation: the test-split ancestral FID is within 0.8 of the training value on MNIST and within 0.4 on CelebA, with the reconstruction arm within 1.2 throughout. Temperature tuning, by contrast, is not a substitute for the refit. Only the mis-scheduled α=0.5 arm benefits from it ($26.60 \to 18.61$ at $\tau = 0.7$ on MNIST); the α=2 arm is best at $\tau = 1$ (11.09, against 13.58 at $\tau = 0.85$ and 24.57 at $\tau = 0.7$), and every CelebA arm degrades under both $\tau = 0.85$ and $\tau = 1.15$.