Uncertainty as an Underconstrained Axis in Lossy Compression

Enzo Tartaglione
LTCI, Télécom Paris, Institut Polytechnique de Paris, France
CNR-ISASI, Italy
enzo.tartaglione@telecom-paris.fr

Abstract

Classical rate-distortion theory minimises expected distortion at a given coding rate but does not constrain the *uncertainty* of the reconstructed signal at the decoder. Predictive uncertainty U is here formalised as the conditional variance of the reconstruction given the source, and shown to be an underconstrained degree of freedom of RD-optimal compression with a measurable downstream effect: at matched rate and matched injected decoder-noise variance, moving a quarter of that variance from the background to the foreground costs up to 25 mIoU points on average in downstream segmentation while pixel-space U stays within 11%, with 89% of (image, quality) pairs showing a significantly negative within-image slope of mIoU on U. A closed-form Uncertainty-Rate-Distortion surface and a single-letter converse for additive-noise channels under a Gaussian assumption with linear regression are derived. A region-importance spatial-URD functional also converts the spatial-allocation finding into a within-image predictor that generalises across two segmentation architectures, three rates, two codecs, and a different segmentation foundation model on a different dataset. Cross-task and cross-modality probes (BLIP captioning, CLIP zero-shot, NYU depth, LibriSpeech speech recognition) show that the spatial-URD prediction transfers to captioning and to speech recognition, and is weak or absent for zero-shot CLIP and monocular depth. Code and data are available at `https://github.com/enzotarta/uncertainty-rate-distortion`.

1 Introduction

Most stored and transmitted images, audio, and video are compressed with loss. The classical theory of Rate-Distortion (RD) [12, 32] prescribes the minimum bit rate R at which a source can be transmitted under a given expected distortion D, and guides the design of codecs for image, video, audio, and speech.

The last decade has shifted compression towards learning-based solutions: neural codecs trained end-to-end on natural data [4, 9, 22] now exceed hand-designed counterparts on standard benchmarks at low and mid bitrates, and enable generative compression for extremely low bitrates [1, 24] and downstream-task-aware coding. They are also increasingly "stochastic", sampling reconstructions from a learned conditional distribution rather than committing to a single estimate, which exposes design degrees of freedom classical RD theory does not address.

The dominant theoretical framework for these stochastic decoders is Rate-Distortion-Perception (R-D-P) [6], which constrains the marginal of the reconstruction. However, this leaves a complementary degree of freedom unspecified: *how much per-source reconstruction variance the decoder commits to at a given rate*. Multiple reconstruction distributions can achieve the same (R, D) while

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

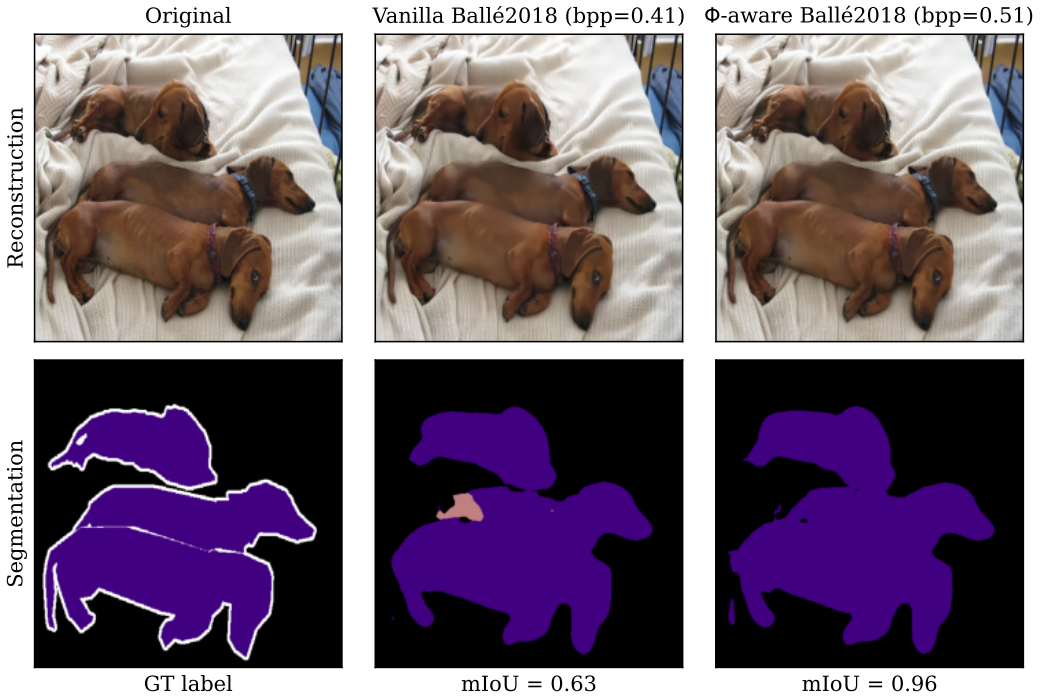

Figure 1: Illustrative example of spatial allocation of reconstruction distortion. Vanilla Ballé2018 [4] at $q = 4$ (centre) vs. a saliency-weighted-MSE fine-tune of the same architecture (right; "Φ-aware" [28]). Both decodings are deterministic ($U = 0$), so this figure shows the allocation of $D(p)$, not of $U(p)$, and the pixel difference is below visual threshold, yet DeepLabV3 flips from mIoU 0.63 (0.41 bpp) to 0.96 (0.51 bpp).

differing substantially in their conditional variance, and where this variance is placed can change the downstream task performance substantially, even when the injected amount is held fixed.

In this work, such a missing degree of freedom is named *predictive uncertainty* U, being the conditional variance of the reconstruction given the source and the bitstream. Its interplay with rate and distortion is here studied, under what is named "the URD framework". A closed-form URD surface for the canonical Gaussian source is derived, and two class-conditional converses are proven, written under a common Cauchy-Schwarz decomposition. Empirically, we observe that at a fixed U, its spatial allocation substantially changes the downstream cost. Besides, the same region-importance principle governs how a fine-tuned codec allocates distortion (Fig. 1), and the same axis is observable in off-the-shelf foundation-model inference via standard sampling-temperature hyperparameters. The contributions are the following:

- A binary-source one-parameter decoder family realising every $U \in [0, \frac{1}{4}]$ at the same RD-optimum (Prop. 1, Sec. 3.1), and a closed-form URD surface for the Gaussian source with a unique non-zero parametric optimum at every rate within the Gaussian test-channel family (Prop. 2, Cor. 1, Sec. 3.2).
- Two class-conditional single-letter converses, written under a common Cauchy-Schwarz decomposition: a two-branch $D \geq D^\star(R, U)$ in the rate-binding regime (lower branch $D \geq U$ in the saturation regime) for additive-noise / linear-mean channels (Th. 1), and $D \geq D_{\mathrm{LM}}(K) + U$ for K-level scalar quantizers via Lloyd-Max (Th. 2); the achievable URD region of the union of these two classes is the lower envelope of the two bowls (Cor. 2, Sec. 3.3).
- A region-importance spatial-URD functional Φ_{fg}. At matched injected variance, mIoU falls monotonically as the share of noise on the foreground grows. Used as a saliency weight on the training distortion of Ballé2018 [4], the same principle gives small rate-matched gains, no better than an MS-SSIM-trained Ballé2018 (Sec. 4.2).

Globally, the contribution is of explanatory nature. URD identifies a degree of freedom that existing objectives leave free and explains why several existing techniques (saliency- and perceptually-weighted losses, sampling temperature) move along it; here, it is not proposed a new state-of-the-art codec.

2 Background and Related Work

Rate-distortion theory and what it leaves unspecified. Shannon's rate-distortion (RD) function $R(D)$ [12, 32] characterises the minimum rate at which a source $X \sim p(x)$ can be coded so that the *expected* distortion $\mathbb{E}[d(X, \hat{X})]$ is below D. For the Gaussian source under MSE, $D(R) = \sigma^2 2^{-2R}$ [5]; modern neural codecs target this trade-off via end-to-end variational training [4, 9, 25]. The RD objective constrains only the *mean* of the per-source distortion: the conditional reconstruction distribution $p(\hat{x} \mid x, b)$ has higher-order statistics (variance, mode count) that classical RD leaves unspecified [12]. Two codecs achieving identical (R, D) may differ in per-image variability of $\hat{X}$, with no RD-level principle for choosing between them. This degree of freedom is the focus of the present paper. We distinguish three forms of uncertainty that are typically conflated.

Aleatoric uncertainty. It consists of irreducible variability inherent in the source itself [20]: in compression, this arises from sensor noise [16], texture randomness [27], or fine-scale variations that cannot be predicted deterministically from coarser representations. Classical RD theory implicitly accommodates aleatoric uncertainty through expected distortion [5, 12] by treating the source as a random variable, but does not distinguish it from other sources of variability in the reconstruction.

Epistemic uncertainty. Unlike the aleatoric one, this consists of incomplete knowledge or limited modelling capacity [17, 20], in principle reducible with more data, expressivity, or supervision. In learned compression systems, it stems from finite training data, restricted model expressiveness [4, 9], or train-test distribution shift, and may also surface in generative-compression setups [2, 7]. Classical compression objectives do not explicitly regulate this form, though it can be addressed with Bayesian or ensemble methods orthogonal to RD design.

Reconstruction-induced uncertainty. A third form, intrinsic to lossy compression itself: even with deterministic source and a perfectly specified model, multiple reconstructions may be consistent with the same compressed representation [20]. This arises from quantization, stochastic decoding (sampling rather than committing to a single $\hat{x}$), or underspecified inverse mappings from compressed representations to signal space [19]. It is distinct from common randomness between encoder and decoder [3], which is a shared resource intentionally exploited; reconstruction-induced uncertainty persists when no such resource is shared. It is also distinct from aleatoric noise in the source and from epistemic uncertainty about the model: it survives perfect source knowledge and optimal coding, and is the degree of freedom that classical RD theory leaves unconstrained. *The body of this paper analyses this third form exclusively*, formalising it as the conditional variance $U = \mathbb{E}[\mathrm{Var}(\hat{X}|X)]$ and characterising its trade-off with R and D.

Related frameworks. The closest framework is the Rate-Distortion-Perception (R-D-P) tradeoff of Blau and Michaeli [6], which extends RD with a perception term $P = d(p_{\hat{X}}, p_X)$ on the *marginal* of $\hat{X}$ and anchors the generative-compression literature [1, 24]. P and U are different objects: P constrains the marginal, U the conditional $p(\hat{x} \mid x)$; the URD-optimal $U^\star$ and the perfect-perception locus $U_{P=0}$ never coincide at $R > 0$ (App. C). They are not independent, however: in the scalar Gaussian test-channel family, fixing R ties U to the reconstruction variance ($\mathrm{Var}(\hat{X}) = U\, 2^{2R}$), and on $U \in [U^\star, U_{P=0}]$ the URD bowl coincides with the Gaussian R-D-P function under W_2 perception [37]. What URD adds in this paper is the class-conditional analysis below and, empirically, the spatial allocation of U, which scalar U does not capture (Sec. 4.2). Prop. 1 shows fixing (R, D) leaves U free across an admissible interval, and Thms. 1, 2 give class-conditional lower bounds $D \geq f(R, U)$. Modern neural codecs based on variational priors [4, 25] and stochastic decoders such as conditional diffusion compression [36] and Cool-Chic [22] expose decoder-side stochasticity but provide no theoretical control over its magnitude, leaving them implicitly on the URD axis without a principle for choosing where. Bayesian deep-learning treatments of uncertainty [17, 20] address parameter

uncertainty θ by marginalising it into a predictive distribution; URD is the reconstruction-induced axis at fixed θ, surviving perfect knowledge of source and model.

3 A URD Perspective

3.1 Preliminaries

Let $X \sim p(x)$ be a source signal and $\hat{X}$ a reconstruction produced from a finite-rate representation. Classical RD theory minimises $R + \lambda D$ with $D = \mathbb{E}[d(X, \hat{X})]$, a constraint on the *expectation* of distortion only, leaving higher-order statistics of $p(\hat{x}|x)$ unconstrained.

Definition 1. *Predictive uncertainty is the per-source variance of the reconstruction over decoder (and any encoder) randomness: $U(x) = \mathrm{Var}(\hat{X} \mid X = x)$, with scalar summary $U = \mathbb{E}[U(X)]$, i.e. the total conditional variance $\mathbb{E}[\mathrm{Var}(\hat{X}|X)]$ used by the theoretical bounds (Sec. 3.2); with a deterministic encoder $U(x)$ equals $\mathrm{Var}(\hat{X} \mid b(x))$. U is defined at the decoder and should not be conflated with entropy models used for coding latent variables (App. C).*

Proposition 1 (Existence of an underconstrained uncertainty axis)**.** *There exist sources and distortion measures for which multiple RD-optimal solutions yield identical (R, D) but different predictive uncertainty.*

Proof. Take $X \in \{0, 1\}$ uniform, ℓ_1 distortion, zero rate. Any distribution supported on $[0, 1]$ achieves the RD optimum $D^\star(0) = 1/2$, since $\frac{1}{2}|\hat{x}| + \frac{1}{2}|1 - \hat{x}|$ equals $\frac{1}{2}$ on $[0, 1]$ and exceeds it outside. The family $\hat{X}_\alpha$ placing mass $\frac{\alpha}{2}$ at each endpoint and $1 - \alpha$ at $\frac{1}{2}$ is supported on $[0, 1]$ for every $\alpha \in [0, 1]$, hence each $\hat{X}_\alpha$ is RD-optimal. The conditional variance $U_\alpha = \alpha/4$ varies continuously over $[0, 1/4]$ as α sweeps $[0, 1]$, so U is not pinned down by RD optimality. □

The reason is that the ℓ_1 RD optimum at zero rate is attained by infinitely many distributions, and U varies freely among them. The same decoupling arises in the practically relevant Gaussian-source / MSE setting when the reconstruction is allowed to be stochastic (Sec. 3.2).

3.2 Gaussian URD Surface

Proposition 2 (URD trade-off for Gaussian sources)**.** *Let $X \sim \mathcal{N}(0, \sigma^2)$, MSE distortion, $U = \mathbb{E}[\mathrm{Var}(\hat{X} \mid X)]$. For Gaussian test channels $\hat{X} = \alpha X + Z$ with $Z \sim \mathcal{N}(0, U)$ independent of X, the minimum achievable distortion at rate $R \geq 0$ and $U \in [0,\ \sigma^2/(2^{2R} - 1)]$ is*

$$D^\star(R, U) \ = \ \left(\sigma - \sqrt{U(2^{2R} - 1)}\right)^2 + U. \tag{1}$$

Corollary 1 (Parametric optimum has non-zero uncertainty)**.** *The minimum of $D^\star(R, U)$ over $U \geq 0$ is uniquely achieved at $U^\star = \sigma^2(2^{2R} - 1)/2^{4R} > 0$, with $D^\star(R, U^\star) = \sigma^2 2^{-2R}$. For any $U \neq U^\star$, $D^\star(R, U) > \sigma^2 2^{-2R}$.*

Fig. 2 shows the bowl at four rates with $U^\star$ marked, and overlays the perfect-perception locus $U_{P=0} = \sigma^2 2^{-2R}$ of R-D-P [6]: $U_{P=0} > U^\star$ at every $R > 0$, so the distortion-optimal and perfect-perception decoders differ. Within this scalar family, the two frameworks are tied: on $U \in [U^\star, U_{P=0}]$ the bowl coincides with the Gaussian R-D-P function under W_2 perception [37] (App. C).

Class-conditional structure. The classical RD function itself is source-and-distortion-class-conditional, with different closed forms across (Gaussian, MSE), (Bernoulli, Hamming), etc. [12]; the achievable URD frontier is analogously channel-class-conditional. Prop. 2's bowl is sharp within the Gaussian test channel family, while K-level Lloyd-Max scalar quantisers (Th. 2) and non-linear posterior-mean decoders (Sec. 4) cut beneath on parts of the U axis. The two operationally relevant classes, additive-noise / linear-mean (Th. 1) and discrete-output (Th. 2), are written under a common Cauchy-Schwarz decomposition (App. A), with their lower envelope the achievable region under the union of these two classes (Cor. 2; an entropy-coded scalar quantiser with more than $\lfloor 2^R \rfloor$ levels can undercut it). Cor. 1's strict $U^\star$ also requires the per-symbol constraint $\mathrm{Var}(\hat{X}|X{=}x){=}U$ a.e.; under the average constraint the left arm collapses by time-sharing (App. A). More generally, $U^\star{>}0$ is a property of the single-letter test-channel family: a deterministic block code reaches $\sigma^2 2^{-2R}$ at $U{=}0$, and the role of shared versus private decoder randomness in such stochastic codes is analysed in [34]. Consistently, every practical codec we probe attains its minimum distortion at $U{=}0$ (Sec. 4).

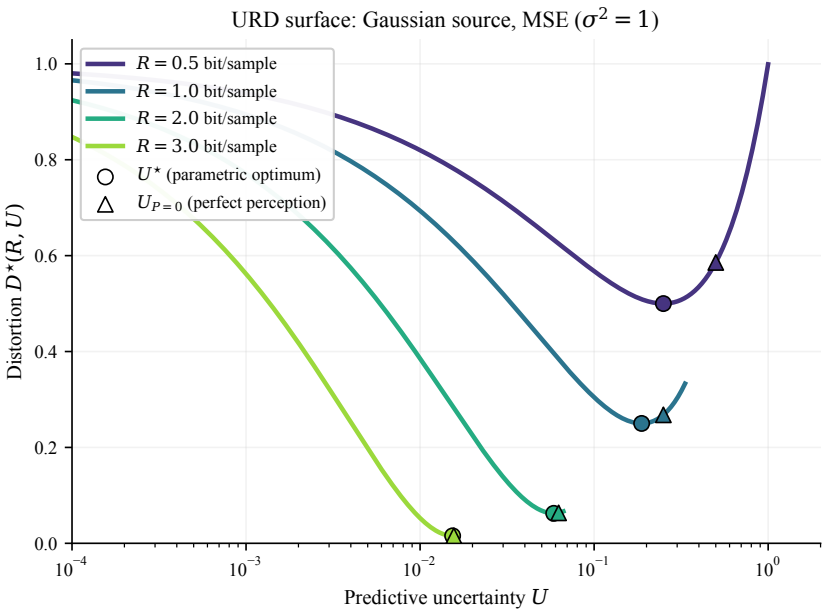

Figure 2: URD trade-off surface $D^\star(R, U)$ for a Gaussian source ($\sigma^2 = 1$) at four coding rates. Filled circles mark the parametric optimum $U^\star$. Open triangles mark the perfect-perception locus $U_{P=0}$ of the Rate-Distortion-Perception framework [6] (App. C): $U_{P=0} > U^\star$ at every rate, so the URD-optimal and perfect-perception decoders never coincide.

3.3 Linear-Regression Converse

Lemma 1 (Gaussian noise minimises mutual information)**.** *Let $\tilde{X} \sim \mathcal{N}(0, V)$, $Z, Z' \perp \tilde{X}$ with $\mathbb{E}[Z] = \mathbb{E}[Z'] = 0$, $\mathrm{Var}(Z) = \mathrm{Var}(Z') = U$, $Z' \sim \mathcal{N}(0, U)$. Then $I(\tilde{X}; \tilde{X} + Z') \le I(\tilde{X}; \tilde{X} + Z)$, with equality iff Z is Gaussian (the worst-additive-noise property [14]).*

Theorem 1 (URD converse for additive-noise / linear-regression channels)**.** *Let $X \sim \mathcal{N}(0, \sigma^2)$ and let $\hat{X} = m(X) + Z$ with $m(X) = \mathbb{E}[\hat{X}|X]$, $Z \perp X$, $\mathbb{E}[Z] = 0$, $\mathrm{Var}(Z) = U$. Under MSE, rate constraint $I(X; \hat{X}) \le R$,* and the linearity assumption $m(x) = \beta x$ *for some $\beta \in [0, 1]$:*

$$D \;\ge\; \begin{cases} D^\star(R,U) = \left(\sigma - \sqrt{U(2^{2R}-1)}\right)^2 + U, & U \in [0,\, \sigma^2/(2^{2R}-1)], \\ U, & U > \sigma^2/(2^{2R}-1), \end{cases} \tag{2}$$

with equality on the upper branch iff $\beta = \alpha^\star = \sqrt{U(2^{2R}-1)}/\sigma$ and $Z \sim \mathcal{N}(0, U)$, and on the lower branch (large-U regime) iff $\beta = 1$ saturates the box constraint.

Proof sketch. With $m(x) = \beta x$ and $Z \perp X$, Lemma 1 gives $|\beta| \le \alpha^\star$ and $D \ge D^\star(R, U)$. Above $U^\star$ the linearity assumption can be dropped: a Gaussian-extremality argument (Th. 3, App. A) extends $D \ge D^\star(R, U)$ to all finite-variance channels for $U \ge U^\star$, with equality iff $\mathbb{E}[\hat{X}|X]$ is linear and $\hat{X}|X$ Gaussian. Below $U^\star$ the universal bound weakens to the Shannon limit $\sigma^2 2^{-2R}$, which deterministic block codes reach at $U = 0$: single-letter K-level Lloyd-Max quantisers cut beneath the bowl on the left arm (64% violation at $R = 1$, $U \to 0$) and posterior-mean decoders cut beneath on natural images (Sec. 4); the discrete-output regime is governed by Th. 2, with achievable URD region of the two classes the lower envelope (Cor. 2).

3.4 Non-Gaussian Sources

Proposition 3 (URD upper envelope for finite-variance sources)**.** *Let X be a zero-mean source with $\sigma^2 < \infty$ and $\hat{X} = \alpha X + Z$ with $Z \sim \mathcal{N}(0, U)$, $Z \perp X$, $\alpha \in [0, 1]$, $U > 0$. Under MSE, $D = (1-\alpha)^2\sigma^2 + U$ and $R(\alpha, U; X) \le \frac{1}{2}\log_2((\alpha^2\sigma^2 + U)/U)$, with equality iff X is Gaussian. Consequently $D^\star_X(R, U) \le D^\star_{\mathrm{Gauss}}(R, U)$ for all $R \ge 0$ and admissible U.*

Within this additive-Gaussian-noise family, the Gaussian bowl upper-bounds the bowl of any finite-variance source under MSE: non-Gaussian sources can only achieve a tighter bowl, never a worse one. As a concrete instance, the Laplacian source $X \sim \mathrm{Laplace}(0, b)$ admits a closed-form Laplace-Gaussian density permitting numerical evaluation of $R(\alpha, U; X)$; the largest relative gap to the Gaussian envelope over U grows from 0.2% at $R = 0.25$ bits to 26% at $R = 2$ bits.

3.5 A Spatial-URD Functional

The scalar U collapses per-position variance into a single number, losing the distinction between "noise on a region the downstream model relies on" and "noise on a region the downstream model ignores." For loss L_f on a deterministic decoding x_{det} under zero-mean noise the first-order Taylor term vanishes in expectation, so the leading contribution to $\mathbb{E}[\Delta L_f]$ is the second-order Hessian trace; with the standard Gauss-Newton diagonal approximation (App. B) this gives the general predictor $\Phi_{w_f}[U] = \int w_f(p)\, U(p)\, dp$ with w_f the Gauss-Newton loss sensitivity; its empirical-Fisher proxy $\|\nabla_x \ell(f(x_{\mathrm{det}}), y_{\mathrm{GT}})\|_p^2$ is dominated on natural images by sparse high-gradient pixels near object boundaries. A binary coarsening replaces this noisy field by a region indicator; for per-pixel classifiers under cross-entropy we take the GT segmentation foreground as that region, a modelling choice motivated (App. B) and tested below. This gives

$$\Phi_{\mathrm{fg}}[U(\cdot)] \;=\; \int_{p \,\in\, \mathrm{fg}} U(p)\, dp, \tag{3}$$

with a deployable U^2-Net [28] saliency proxy in place of the GT mask (Sec. 4.2). Φ_{fg} is the segmentation instance of Φ_{w_f} with w_f replaced by the indicator of a task region $M_f^\star$; cross-task lifts and its limits are tested in Sec. 4.2.

4 Empirical Study

Throughout, Ballé2018 denotes the factorized-prior model of [4] (CompressAI `bmshj2018-factorized`). Protocols and full results that do not fit here are in the code release. The right-arm prediction of Prop. 2 (D rises monotonically in U at every rate) is consistent with Ballé2018 + Kodak [4, 21] across six quality levels: PSNR is rate-ordered at low U and collapses to a common right-arm at $U \gtrsim 10^{-2}$ (within 1 dB), with negligible U already costing several dB; here $D \approx D_{\mathrm{det}} + U$, which lies below the bowl calibrated to the codec's deterministic distortion at small U and above it at larger U. The rest of this section studies the image-by-image downstream cost of decoder uncertainty (Sec. 4.1), spatial allocation as the dominant axis (Sec. 4.2), and the URD axis in deployed foundation models (Sec. 4.3).

4.1 Decoder uncertainty has a measurable downstream cost

We evaluate semantic segmentation on Pascal VOC 2012 [15] with DeepLabV3-ResNet101 [8] pretrained on COCO+VOC. For each of 150 validation images, the Ballé2018 [4] latent is fixed, decoder noise of standard deviation $\tau \in \{0, 0.1, 0.3, 0.8, 1.6\}$ is injected, $K = 15$ stochastic reconstructions per τ are passed through the segmentation network, and per-image mIoU is recorded. Within-image regression of mIoU on U across all noise levels (the deterministic decode included) produces a strongly negative slope distribution (Fig. 3): median slopes $-15.4, -22.2, -9.7$ at $q = 2, 4, 6$ respectively, with 89% (pooled; 83-93% per rate) of slopes significantly negative (one-sided $p < 0.05$). Wilcoxon signed-rank test on the pooled 450 slopes rejects the null at $p < 10^{-65}$. Mean mIoU loss at $q = 2$ is -0.150 (95% bootstrap CI strictly below zero). *Adding decoder uncertainty at a fixed rate (constant bitstream, latent perturbed) monotonically and significantly degrades downstream segmentation performance.* Because the injected noise is independent of the deterministic decode, $D \approx D_{\mathrm{det}} + U$ here, so this experiment does not separate the cost of U from that of D: regressing mIoU jointly on (D, P, U) gives no negative coefficient on U, and within $D \times P$ bins the slope on U is not significant. Sec. 4.2 holds the injected variance fixed and varies only its allocation.

4.2 Spatial allocation of U governs downstream cost

The theory of Sec. 3 fixes the best *scalar* operating point. Here the total injected variance is fixed and only *where* it sits changes, which scalar U, having no spatial structure, cannot describe.

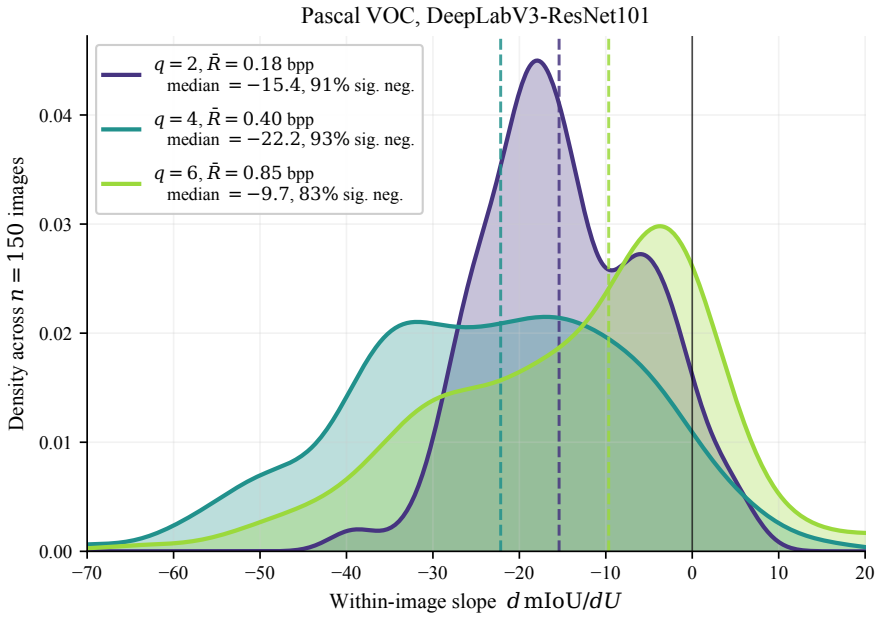

Figure 3: Pascal VOC 2012 segmentation [8, 15], 150 images, Ballé2018 [4] at $q \in \{2, 4, 6\}$, DeepLabV3-ResNet101. Per-image within-image regression slopes $d\,\mathrm{mIoU}/dU$ are concentrated in the negative half-line at every rate; 83-93% of slopes are significantly negative ($p < 0.05$). Vertical dashed lines mark per-rate medians.

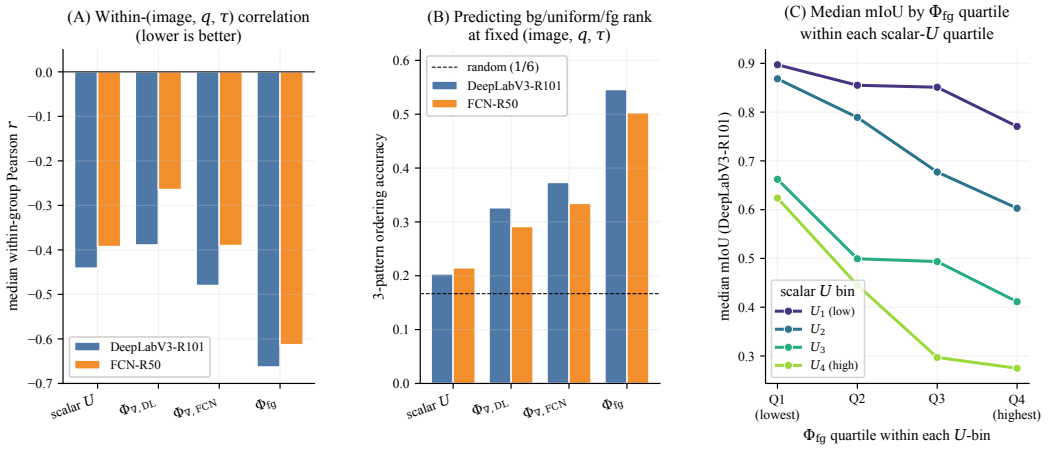

Figure 4: Spatial-URD result on the Pascal VOC isolation experiment (300 images, $q \in \{2, 4, 6\}$, two segmentation models, $n = 2700$ triples). **(A)** Median Pearson r between mIoU and each predictor within (image, q, τ) groups. **(B)** 3-pattern (bg/uniform/fg) ordering accuracy. **(C)** Matched-U stratification: within each scalar-U quartile, Φ_{fg} further partitions median DL mIoU monotonically.

Setup. For each of 300 Pascal VOC images, Ballé2018 [4] at $q \in \{2, 4, 6\}$, three nominal noise scales $\tau_{\mathrm{base}} \in \{0.4, 0.8, 1.6\}$, decoder-side latent noise is injected under three spatial allocations: *uniform* (everywhere), *fg-only* (latent positions whose receptive field overlaps the GT foreground), and *bg-only* (the complementary mask). The per-position scale is $\tau_{\mathrm{base}}/\sqrt{p}$, where p is the active-mask fraction, holding total injected latent variance constant. Partial patterns interpolate between the two: weights r and $\sqrt{1 - r^2}$ on independent fg-only and bg-only fields put a fraction r^2 of the injected variance on the foreground, $r \in \{0.25, 0.5, 0.75\}$, at the same total. Ten stochastic decodes per (image, pattern, τ)

yield per-sample D, P, scalar U, and mIoU; two segmentation models (DeepLabV3-ResNet101 and FCN-ResNet50, both COCO+VOC pretrained) are evaluated.

Scalar U is insufficient. Total injected latent variance is held constant across allocations, so the same image at the same rate yields reconstructions with very different downstream cost. Pixel-space U is not exactly matched: the decoder amplifies concentrated noise, so fg-only reconstructions carry a median (over images) 1.6-1.8$\times$ the pixel-space U of bg-only, while uniform carries slightly less (0.93-0.95$\times$). The most closely matched comparison is the partial family: from bg-only to $r=0.5$ (a quarter of the variance on the foreground), pixel-space U stays within 0.89-1.04$\times$ of bg-only, yet mean DL mIoU drops by 13-25% at $\tau_{\text{base}}=1.6$, and it decreases monotonically along bg-only $\to r=0.25 \to 0.5 \to 0.75 \to$ fg-only in all nine (q, τ) settings. At $\tau_{\text{base}}=1.6$ and $q=4$, swapping uniform for fg-only costs $\Delta\text{mIoU}_{\text{DL}}=-0.109$ (Wilcoxon $p<10^{-17}$); swapping uniform for bg-only *gains* $+0.253$ ($p<10^{-42}$); uniform and bg-only are the near-matched pair. On an illustrative image, fg-only and bg-only at matched injected variance (pixel-space U 0.022 vs. 0.017) yield mIoU 0.21 vs. 0.99, a 78% gap within-image swing; this image is near the top of the distribution (98th percentile), and the median swing over the 300 images is 41%.

Validating the predictors of Sec. 3.5. Median Pearson r within (image, q, τ) groups: Φ_{fg} at -0.66 (DL) / -0.61 (FCN), versus the gradient-based $\Phi_{\nabla,f}$ at -0.39/-0.39 and scalar U at -0.44/-0.39 (Eq. 3, Fig. 4 left). 3-pattern ordering accuracy across 2700 triples: Φ_{fg} at 54.6%/50.3%, $\Phi_{\nabla,f}$ at 32.6%/33.4%, scalar U at 20.3%/21.5% (chance 16.7%); accuracy is stable across rates (55.4%, 55.4%, 52.9% at $q=2,4,6$ for DL). Because the patterns are built from the same foreground mask, Φ_{fg} predicts the order bg-only $\succ$ uniform $\succ$ fg-only in 99.8% of triples; its accuracy is therefore the frequency with which that damage order is realised, and the informative contrast is with scalar U, which stays close to chance. Adding the gradient saliency on top of the foreground mask gives almost no gain (54.7% vs. 54.6% for DL), consistent with the coarsening argument of App. B: the foreground mask alone carries the signal. Within each quartile of scalar U, Φ_{fg} further partitions median mIoU monotonically (Fig. 4, right): in the highest-U bin, Φ_{fg}-Q1 vs. Q4 differ by 0.35 in median DL mIoU at matched scalar U.

Cityscapes robustness check (different dataset and architecture). We re-run the spatial-isolation pipeline with SAM2.1-Hiera-Tiny [31] on 300 Cityscapes val [11] images (279 usable), prompted with the bbox of the single largest annotated instance and using per-instance IoU between SAM2's mask and the GT instance mask as the metric. Mean ΔIoU vs the deterministic baseline at $q=2, \tau=1.6$ is bg-only -0.022, fg-only -0.165, uniform -0.224: *any* allocation that touches the foreground damages IoU substantially while bg-only barely changes it, the qualitative Φ_{fg} prediction. The strict bg<uniform<fg damage ranking that Φ_{fg} would imply does not hold at the population mean (uniform is worse than fg-only here), but it is what Φ_{fg} is designed to predict per-image triple: ordering accuracy across 2,637 triples is 52.5% versus 13.1% for scalar U (chance 16.7%), with median within-(image, q, τ) $r=-0.63$ within 0.03 of DL-R101 on VOC (54.6%, $r=-0.66$), so the framework transfers across both architecture (from semantic to instance segmentation) and dataset (from VOC objects to Cityscapes street scenes). The same SAM2 model fed loose multi-class bbox prompts on VOC produces 21.4%/sign-reversed r, consistent with the scope condition that the prompt must approximately match the task region (Fig. 5). Φ_{fg}'s predictive power therefore depends on the mask (GT or estimated saliency) matching the region the downstream model actually uses.

Scope. The result transfers across codecs (cheng2020 [9] 52.4%/47.1%), to a deployable U^2-Net saliency proxy, and cross-task to BLIP [23] captioning and Whisper-Encodec ASR; it is weak or absent for DepthAnything-V2 [35] depth and CLIP [29] zero-shot classification, which we read as a scope limit of Φ_{fg}.

A Φ_{fg}-aware codec. As a complementary test for whether the same spatial-allocation principle *emerges from gradient descent*, we fine-tune Ballé2018 [4] ($q \in \{2,3,4,5\}$) on VOC train with $\mathcal{L}_\Phi = \lambda \cdot 255^2 \cdot \mathbb{E}_p[(1+\alpha s(p))(x(p)-\hat{x}(p))^2] + \text{bpp}$, s an off-the-shelf U^2-Net saliency map. At $\alpha=2$, against a densified vanilla curve (four extra intermediate-rate checkpoints), the rate-matched change is FCN $\Delta \in [-0.005, +0.015]$ and DL $\Delta \in [+0.001, +0.009]$, significant (paired Wilcoxon $p<0.05$) for FCN at 3 of 4 rates and for DL at 1 of 4. In a separate α-sweep at $q=4$ against the five pretrained vanilla points, $\alpha=2$ gains $+0.009$ DL / $+0.008$ FCN and the saliency-blind $\alpha=0$ fine-tune about half as much (not significant), but $\alpha=2$ is not significantly better than $\alpha=0$ either, so the experiment does not isolate the saliency weight. MS-SSIM-trained Ballé is equal or slightly better at every rate where both are available, except DL at $q=4$ (0.782 vs. 0.783). We read this as a

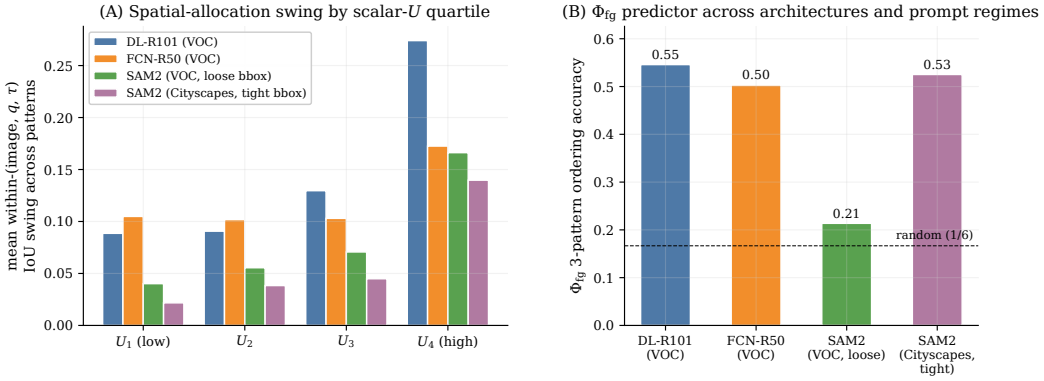

Figure 5: Φ_{fg} across architectures and datasets. **(A)** Within-(image, q, τ) IoU swing across the three spatial allocations, by scalar-U quartile. **(B)** Φ_{fg} 3-pattern ordering accuracy: tight-prompt SAM2 on Cityscapes (52.5%) is within 0.03 of DL-R101 on VOC (54.6%); loose-prompt SAM2 on VOC (21.4%) collapses (the prompt-matching scope condition of Sec. 4.2).

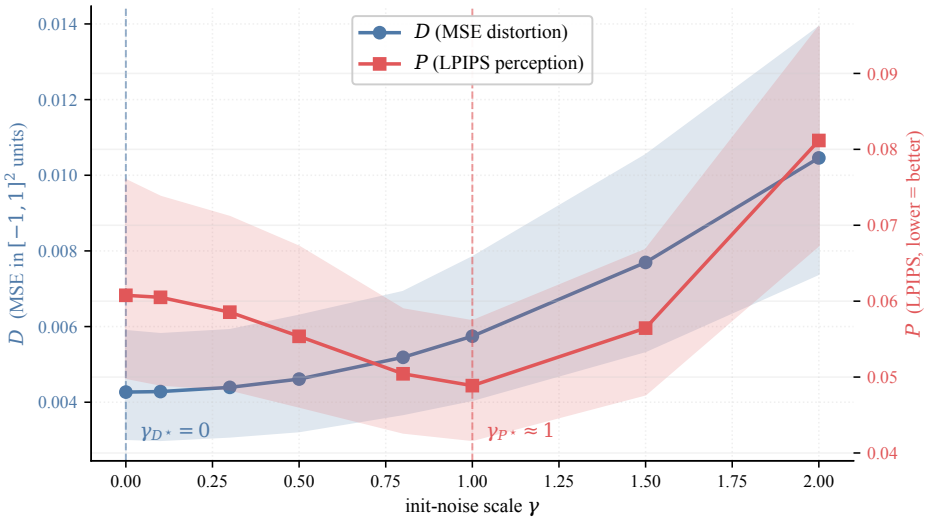

Figure 6: MSE and LPIPS are minimised at different points of the decoder-noise scale γ of a deployed diffusion codec (CDC [36], LPIPS-trained, $\bar{R} = 0.70$ bpp on Kodak); shaded: 95% bootstrap CIs over the 24 images.

consistency check rather than a new codec objective: the codec is deterministic ($U = 0$), so it tests the region-importance principle on the allocation of distortion, not of U.

4.3 URD in deployed foundation models

The codec experiments probe URD via constructed decoder-side perturbations. We complement them with off-the-shelf demonstrations that the axis exists: only the sampling temperature T is varied on two deployed foundation models, $K = 4$ decodes per (input, T); the larger is the Encodec+Whisper sweep (202 utterances). U is the mean per-token Shannon entropy of the T-scaled categorical.

ASR (Whisper-tiny.en + LibriSpeech [30, 26], 30 utts). Sweeping $T \in \{0, 0.3, 0.7, 1.0, 1.5\}$, WER is approximately flat at ~ 0.07 for $T \leq 0.7$ and rises monotonically to **0.78** at $T = 1.5$ ($\sim$10.7$\times$ degradation, the URD right-arm prediction), while U_{audio} spans more than two orders of magnitude on $T \in [0.3, 1.5]$ (from 0.021 to 4.09 bits/tok); within-(utt, T) median $r(\text{WER}, U_{\text{audio}}) = +0.85$ (95% CI $[+0.68, +0.91]$, 67/83 groups positive, sign test $p = 7 \times 10^{-9}$).

LLM reasoning (Qwen2.5-1.5B-Instruct [33] + GSM8K [10], 30 problems). Sweeping $T \in \{0.1, 0.5, 1.0, 1.5\}$, mean accuracy declines gently (0.63, 0.60, 0.50) for $T \leq 1$ and collapses to $\mathbf{0.02}$ at $T = 1.5$ (2 of 120 generations correct) as mean U_{tok} grows from 0.026 to $\mathbf{14.9}$ bits/tok (~ 2.8 orders of magnitude). The matched-U per-token-Φ analogue requires explicit per-token temperature schedules (Sec. 4.2's design) and is not isolable at $K = 4$ uniform-T samples; we report it only as evidence that the axis exists for LLMs, the collapse at $T = 1.5$ being the only clear effect at this sample size.

Cross-modal temporal-Φ (Encodec [13] + Whisper, 202 utts). At matched scalar latent variance with a VAD-derived voice mask, mean ΔWER at $\tau = 1.6$ is voice-only $+0.097$, uniform $+0.067$, silence-only $+0.011$: noise on voice frames damages WER $\sim 9\times$ more than on silence (paired Wilcoxon $p < 10^{-24}$; silence-only is best on 82.7% of 606 triples vs. 33.3% chance). The spatial-allocation principle of Sec. 4.2 transfers cross-modality.

Diffusion decoders. CDC [36] decodes by DDIM iteration from $x_T = \gamma\varepsilon$; the MSE-distortion-optimal $\gamma_{D^\star} = 0$ and LPIPS-optimal $\gamma_{P^\star} \approx 1$ are different points on the same URD axis (Fig. 6). This is the familiar tension between distortion and perceptual quality (LPIPS being a full-reference proxy for the latter, not the marginal divergence of R-D-P), observed here along the decoder-noise scale of a deployed codec with no constructed perturbation.

Limitations and scope. The closed-form URD surface and the right-arm extremality argument assume a Gaussian source under MSE; the finite-variance upper envelope (Prop. 3) and a numerical Laplacian benchmark cover other sources, but source-specific converses remain open, as does the lower branch of Th. 1 (still linear-mean). The two-bowl envelope (Cor. 2) is the achievable region of its two classes only: entropy-coded scalar quantisers can undercut it, and a converse over all discrete-output channels is open. $U^\star > 0$ is a property of the test-channel family; the practical codecs we probe attain their minimum distortion at $U = 0$. Empirically, Φ_{fg} is task-class-specific: designed for per-pixel cross-entropy classifiers, it is weak or absent on CLIP and DepthAnything-V2 and on SAM2 collapses when the prompt does not approximately match the task region. The codec validation covers Ballé2018 [4] / cheng2020 [9]; transformer and diffusion-prior codecs are open. Finally, the foundation-model probes only show that the axis exists; matched-U per-token-Φ schedules isolable at small K are future work.

5 Conclusion

Classical rate-distortion leaves a degree of freedom unspecified: the predictive uncertainty U a decoder commits to per source. On the Gaussian source, the URD surface is a closed form, two class-conditional converses fall out of a common Cauchy-Schwarz decomposition, and a Gaussian-extremality argument proves the bowl's right arm universal across finite-variance channels. Empirically, at matched injected variance, downstream mIoU falls monotonically as the share of decoder noise on the foreground grows. The functional Φ_{fg} predicts within-image mIoU rankings on more than half of Pascal VOC triples and transfers to Cityscapes; as a saliency weight on the training distortion it gives small rate-matched gains; the same axis surfaces off-the-shelf on Whisper and on a deployed diffusion codec where MSE and LPIPS sit at different optima along the same noise scale.

On the theory side, source-specific converses for non-Gaussian sources, a discrete-output bound for entropy-coded scalar quantisers, and a function-valued URD surface where $U(\cdot)$ is a field rather than a scalar are the natural next objects. On the engineering side, the spatial-allocation principle generalises beyond saliency-weighted MSE to any classifier-derived Hessian, suggesting a family of Φ_f-aware codecs trained without task labels; per-token-Φ schedules at variable T are the next experiment for autoregressive decoders.

Acknowledgments and Disclosure of Funding

This work benefited from Hi! PARIS and State funding managed by the French National Research Agency (ANR) under the France 2030 program, reference ANR-23-IACL-0005. This work was also supported by the Italian MUR through the FIS 2 project FIS-2023-03417 (DD n. 23314 10-12-2024, CUP D53C25002040001).

References

[1] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[2] Tom Bachard, Tom Bordin, and Thomas Maugey. Coclico: Extremely low bitrate image compression based on clip semantic and tiny color map. In *2024 Picture Coding Symposium (PCS)*. IEEE, 2024.

[3] Jona Ballé, Luca Versari, Emilien Dupont, Hyunjik Kim, and Matthias Bauer. Good, cheap, and fast: Overfitted image compression with wasserstein distortion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

[4] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

[5] Toby Berger. Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*, 2003.

[6] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*. PMLR, 2019.

[7] Tom Bordin and Thomas Maugey. Semantic based generative compression of images for extremely low bitrates. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2023.

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[9] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[13] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023.

[14] Suhas N. Diggavi and Thomas M. Cover. The worst additive noise under a covariance constraint. *IEEE Transactions on Information Theory*, 47(7):3072–3081, 2001.

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 2010.

[16] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE transactions on image processing*, 17, 2008.

[17] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 2016.

[18] Allen Gersho and Robert M Gray. *Vector quantization and signal compression*. Springer Science & Business Media, 2012.

[19] Jari P Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*. Springer, 2005.

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[21] Eastman Kodak. Kodak lossless true color image suite (photocd pcd0992), 1993. URL `http://r0k.us/graphics/kodak/`.

[22] Théo Ladune, Pierrick Philippe, Félix Henry, Gordon Clare, and Thomas Leguay. Cool-chic: Coordinate-based low complexity hierarchical image codec. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023.

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 2022.

[24] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in neural information processing systems*, 33, 2020.

[25] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018.

[26] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015.

[27] Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40, 2000.

[28] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106, 2020.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 2021.

[30] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 2023.

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, 2025.

[32] Claude E Shannon. Coding theorems for a discrete source with a fidelity criterion, 1993.

[33] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[34] Aaron B. Wagner. The rate-distortion-perception trade-off: The role of common randomness. *arXiv preprint arXiv:2202.04147*, 2022.

[35] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37, 2024.

[36] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.

[37] George Zhang, Jingjing Qian, Jun Chen, and Ashish Khisti. Universal rate-distortion-perception representations for lossy compression. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

A Proofs

Proposition 1 is proved in the main text, and Lemma 1 is the classical worst-additive-noise result [14]; the remaining proofs follow.

Proof of Proposition 2. For the test channel $\hat{X}|X \sim \mathcal{N}(\alpha X, U)$ with $\mathrm{Var}(Z) = U$,

$$R \;=\; \tfrac{1}{2}\log_2 \tfrac{\alpha^2\sigma^2+U}{U}, \qquad D \;=\; (1-\alpha)^2\sigma^2 + U.$$

The rate constraint fixes $\alpha \le \alpha^\star := \sqrt{U(2^{2R}-1)/\sigma^2}$. Since D is strictly decreasing in α on $[0,1]$, the minimum at $\alpha = \alpha^\star$ gives

$$D^\star(R,U) \;=\; (\sigma - \sqrt{U(2^{2R}-1)})^2 + U. \qquad \square$$

Proof of Corollary 1. Set $\partial D^\star/\partial U = 0$:

$$-\sigma\sqrt{(2^{2R}-1)/U} + 2^{2R} \;=\; 0 \quad\Longrightarrow\quad U^\star \;=\; \sigma^2(2^{2R}-1)/2^{4R}.$$

Substituting back yields $D^\star(R,U^\star) = \sigma^2 2^{-2R}$. Strict convexity of $D^\star$ in $\sqrt{U}$ gives uniqueness. $\square$

Proof of Theorem 1. Idea. Linearity of m and the Gaussian source make the channel a Gaussian linear channel up to noise shape; Lemma 1 then forces the noise to be the worst-case Gaussian, and the rate constraint caps β at $\alpha^\star$.

Step 1 (distortion identity). With $m(x) = \beta x$ and $Z \perp X$, $\mathbb{E}[Z] = 0$, $\mathrm{Var}(Z) = U$:

$$D \;=\; \mathbb{E}[(X - \beta X - Z)^2] \;=\; (1-\beta)^2\sigma^2 + U.$$

Step 2 (rate constraint via Lemma 1). Apply Lemma 1 with $\tilde{X} = \beta X \sim \mathcal{N}(0, \beta^2\sigma^2)$ and $Z' \sim \mathcal{N}(0,U)$:

$$I(X;\beta X + Z') \;\le\; I(X;\beta X + Z) \;\le\; R.$$

The Gaussian-conditional channel evaluates to $I(X;\beta X + Z') = \frac{1}{2}\log_2(1 + \beta^2\sigma^2/U)$. Combining,

$$\beta^2\sigma^2/U \;\le\; 2^{2R} - 1 \quad\Longleftrightarrow\quad |\beta| \;\le\; \alpha^\star.$$

Step 3 (minimise D over the feasible set). $(1-\beta)^2$ is strictly decreasing in β on $[0,1]$, so $D \ge (1-\alpha^\star)^2\sigma^2 + U = D^\star(R,U)$. Equality holds iff $\beta = \alpha^\star$ and $Z \sim \mathcal{N}(0,U)$. $\square$

Time-sharing degeneracy under the average constraint. Under $\mathbb{E}[\mathrm{Var}(\hat{X} \mid X)] \le U$, a Shannon-optimal deterministic codec achieves $(R, D = \sigma^2 2^{-2R}, U = 0)$, satisfying the average constraint trivially for any $U \ge 0$. Time-sharing it with the Gaussian test channel at $U^\star$ yields $(R, D = \sigma^2 2^{-2R}, U)$ for every $U \in [0, U^\star]$, so the achievable URD frontier is flat at $D = \sigma^2 2^{-2R}$ across the entire left arm; Cor. 1's strict optimality is recovered by the per-symbol equality constraint $\mathrm{Var}(\hat{X} \mid X = x) = U$ for almost every x. Mixing two channels with conditional means m_A, m_B produces conditional variance $U + \lambda(1-\lambda)(m_A(x) - m_B(x))^2 > U$ wherever the means differ, so per-symbol equality rules out time-sharing of distinct channels and corresponds to the natural physical setting where the decoder injects a single calibrated source of randomness rather than alternating between operating modes. Per-symbol equality does not, however, exclude deterministic codes: a deterministic block code satisfies it with $U = 0$ and reaches $\sigma^2 2^{-2R}$ as the block length grows. Cor. 1 is therefore a statement about the single-letter test-channel family; the role of shared versus private randomness in stochastic codes is analysed in [34].

A common Cauchy-Schwarz decomposition. Theorem 1 does not extend to non-linear m: the one-bit Lloyd-Max channel $m(X) = \sigma\sqrt{2/\pi}\,\mathrm{sign}(X)$ plus noise reaches $D = \sigma^2(1 - 2/\pi) + U$ at $R \le 1$ bit, below the bowl for small U. For any zero-mean source X with $\sigma^2 = \mathrm{Var}(X) < \infty$ and any channel $\hat{X} = m(X) + Z$ with $Z \perp X$, $\mathbb{E}[Z] = 0$, $\mathrm{Var}(Z) = U$,

$$D \;=\; \mathbb{E}[(X-m(X))^2] + U = \sigma^2 - 2\rho\sigma\sqrt{V_M} + V_M + U$$
$$= \left(\sqrt{V_M} - \rho\sigma\right)^2 \;+\; \sigma^2(1-\rho^2) \;+\; U, \tag{4}$$

where $V_M := \mathbb{E}[m(X)^2]$ and $\rho := \mathbb{E}[Xm(X)]/(\sigma\sqrt{V_M}) \in [-1,1]$ is the correlation. The first term is non-negative (zero iff $\sqrt{V_M} = \rho\sigma$, i.e. the channel scales optimally for its ρ); the second is non-negative (zero iff $|\rho| = 1$, the linear case). Hence the universal lower bound

$$D \;\geq\; \sigma^2(1-\rho^2) + U. \tag{5}$$

For linear m ($\rho{=}1$) the rate constraint caps V_M and the first term gives Theorem 1; for K-level m the rate constraint does not bind V_M and the second term gives Theorem 2.

Theorem 2 (Discrete-output (atomic) URD converse)**.** *Let $X \sim \mathcal{N}(0,\sigma^2)$ and let $\hat{X} = m(X) + Z$ with $m : \mathbb{R} \to \mathcal{L}$ for some $\mathcal{L} \subset \mathbb{R}$ with $|\mathcal{L}| \leq K$, $Z \perp X$, $\mathbb{E}[Z] = 0$, $\mathrm{Var}(Z) = U$. Then*

$$D \;\geq\; D_{\mathrm{LM}}(K) + U, \qquad D_{\mathrm{LM}}(K) := \sigma^2\big(1 - \rho^2_{\mathrm{LM}}(K)\big), \tag{6}$$

where $\rho_{\mathrm{LM}}(K) = \mathrm{Corr}(X, m_K^{\mathrm{LM}}(X))$ is the correlation achieved by the K-level Lloyd-Max quantiser of the standard Gaussian. Equality holds iff m coincides with the K-level Lloyd-Max minimiser (Voronoi-cell encoding with optimal levels). Combined with a rate constraint $I(X;\hat{X}) \leq R$, the bound applies for every K such that the K-level channel achieves $I \leq R$; for integer R and $K = 2^R$ this is automatic.

Proof of Theorem 2. Idea. The atomic constraint reduces the URD problem to deterministic K-level quantisation plus an additive-noise tax of U; the deterministic piece is solved by Lloyd-Max.

Step 1 (decompose). Using $Z \perp X$, $\mathbb{E}[Z] = 0$, $\mathrm{Var}(Z) = U$ in (4),

$$D \;=\; \mathbb{E}[(X-m(X))^2] \;+\; U.$$

The first term is the deterministic-quantisation MSE of m.

Step 2 (Lloyd-Max optimality). Minimising $\mathbb{E}[(X-m(X))^2]$ over $m : \mathbb{R} \to \mathcal{L}$ with $|\mathcal{L}| \leq K$ is the textbook K-level Lloyd-Max problem on the standard Gaussian; the minimum is $D_{\mathrm{LM}}(K)$, achieved (up to relabelling) by the optimal Lloyd-Max levels with Voronoi-cell encoding [18]. Substitution yields (6).

Step 3 (rate-constraint refinement). The K-level channel input has $H(m(X)) \leq \log_2 K$, so $I(X;\hat{X}) \leq H(m(X)) \leq \log_2 K \leq R$ whenever $K \leq 2^R$. □

Corollary 2 (Lower envelope of class-conditional bounds)**.** *For Gaussian source $X \sim \mathcal{N}(0,\sigma^2)$ at rate $R \geq 0$ and uncertainty $U \geq 0$, the achievable distortion under the union of (i) the linear-regression Gaussian-noise channel class (Th. 1) and (ii) the K-level discrete-output channel class with $K \leq 2^R$ (Th. 2) satisfies*

$$D_{\min}(R,U) \;\geq\; \min\Big(\underbrace{D^{\star}_{\mathrm{lin}}(R,U)}_{\text{Th. 1, floored}},\; \underbrace{D_{\mathrm{LM}}(\lfloor 2^R \rfloor) + U}_{\text{Th. 2}}\Big),$$
$$D^{\star}_{\mathrm{lin}}(R,U) = \begin{cases} (\sigma - \sqrt{U(2^{2R}-1)})^2 + U, & U \leq \sigma^2/(2^{2R}-1), \\ U, & U > \sigma^2/(2^{2R}-1). \end{cases} \tag{7}$$

Each side is achievable within its respective class. The crossover where the two bounds intersect partitions the (U,R) plane into a regime where the linear Gaussian channel is preferred (large U or large R, where the rate constraint binds) and a regime where Lloyd-Max-style scalar quantisation is preferred (small U or small R, where the input alphabet is the binding constraint). At $U \geq \sigma^2/(2^{2R}-1)$, the linear branch reduces to the trivial bound $D \geq U$ (achieved by $\beta{=}1$ saturation), so the linear branch attains the envelope there.

Corollary 2 is the achievable region of these two classes only: entropy-coded scalar quantisers with more than $\lfloor 2^R \rfloor$ levels can undercut it, at integer rates too (at $R{=}1$ bit, a symmetric 3-level quantiser with thresholds $\pm 1.21\sigma$ has entropy 0.998 bits and distortion $0.349\sigma^2 < D_{\mathrm{LM}}(2){=}0.363\sigma^2$).

Theorem 3 (Right-arm universality on Gaussian source)**.** *Let $X \sim \mathcal{N}(0, \sigma^2)$ and let $\hat{X}$ be any random variable with finite variance and $\mathbb{E}[\hat{X}] = 0$, jointly distributed with X such that $I(X; \hat{X}) \le R$. Define $V_M := \mathrm{Var}(\mathbb{E}[\hat{X}|X])$ and $U := \mathbb{E}[\mathrm{Var}(\hat{X}|X)]$. Under MSE distortion: for $U \in [U^\star,\ \sigma^2/(2^{2R}-1)]$,*

$$D \ \ge \ D^\star(R, U) = \big(\sigma - \sqrt{U(2^{2R}-1)}\big)^2 + U, \tag{8}$$

with equality iff $\mathbb{E}[\hat{X}|X]$ is linear in X and $\hat{X}|X$ is Gaussian. For $U > \sigma^2/(2^{2R}-1)$, $D \ge U$ (saturation); for $U < U^\star$, $D \ge \sigma^2 \cdot 2^{-2R}$ (the Shannon RD limit, looser than the linear-class $D^\star(R, U)$ on the left arm).

Proof of Theorem 3. Idea. Decompose the MSE by the law of total variance into a linear-projection piece, a non-linearity penalty, and the noise variance U. Two constraints bind the linear-projection coefficient a_1: a Cauchy-Schwarz cap (always tight under linearity) and a Gaussian-extremality cap from the rate budget. The optimum partitions into two cases by which constraint binds.

Step 1 (MSE decomposition). Let $a_1 := \mathrm{Cov}(X, \hat{X})/\sigma$ and $V_{\hat{X}} := \mathrm{Var}(\hat{X}) = V_M + U$ (law of total variance). Then

$$D \ = \ \sigma^2 - 2\,\mathrm{Cov}(X, \hat{X}) + V_{\hat{X}} \ = \ (\sigma - a_1)^2 + (V_M - a_1^2) + U.$$

Step 2 (two constraints on (a_1, V_M)).

(i) **Cauchy-Schwarz on the conditional mean.** $\mathrm{Cov}(X, \hat{X}) = \mathrm{Cov}(X, \mathbb{E}[\hat{X}|X])$, so $|a_1\sigma| \le \sigma\sqrt{V_M}$, i.e. $a_1^2 \le V_M$.

(ii) **Gaussian extremality of MI.** For X Gaussian and any joint $(X, \hat{X})$, the variance-and-correlation-matched Gaussian replacement $\tilde{X}$ satisfies $I(X; \tilde{X}) \le I(X; \hat{X})$ [12]. The Gaussian channel evaluates to $I(X; \tilde{X}) = -\frac{1}{2}\log_2(1 - \rho^2)$, where $\rho = a_1/\sqrt{V_{\hat{X}}}$. The rate budget $I(X; \hat{X}) \le R$ then yields

$$\rho^2 \le 1 - 2^{-2R} \quad \Longleftrightarrow \quad a_1^2 \le (1 - 2^{-2R})(V_M + U).$$

Step 3 (case split). Minimise $D = (\sigma - a_1)^2 + (V_M - a_1^2) + U$ subject to (i) and (ii). For fixed a_1, the optimal V_M saturates whichever constraint binds:

$$V_M^\star \ = \ \max\big(a_1^2,\ a_1^2/(1 - 2^{-2R}) - U\big).$$

Case A: (i) binds, $a_1^2 \le U(2^{2R}-1)$. Then $V_M = a_1^2$ and $D = (\sigma - a_1)^2 + U$. Optimal $a_1 = \min(\sigma, \sqrt{U(2^{2R}-1)})$:

$$U \le \sigma^2/(2^{2R}-1):\ D = D^\star(R, U); \quad U > \sigma^2/(2^{2R}-1):\ D = U.$$

Case B: (ii) binds, $a_1^2 > U(2^{2R}-1)$. Then $V_M = a_1^2/(1 - 2^{-2R}) - U$ and $D = (\sigma - a_1)^2 + a_1^2/(2^{2R}-1)$. Optimal $a_1 = \sigma(2^{2R}-1)/2^{2R}$ gives $D = \sigma^2 \cdot 2^{-2R}$. Feasibility ($a_1^2 > U(2^{2R}-1)$) requires $U < U^\star$.

Step 4 (combine the two cases).

- For $U \le U^\star$, Case B's optimum is feasible and yields the Shannon limit $\sigma^2 \cdot 2^{-2R}$.
- For $U^\star \le U \le \sigma^2/(2^{2R}-1)$, Case B is infeasible, so Case A binds: $D \ge D^\star(R, U)$.
- For $U > \sigma^2/(2^{2R}-1)$, only the saturation branch remains: $D \ge U$.

Equality on the right arm ($U^\star \le U \le \sigma^2/(2^{2R}-1)$) requires (i) and (ii) tight simultaneously, i.e. $\mathbb{E}[\hat{X}|X]$ linear in X and $\hat{X}|X$ Gaussian. □

Proof of Proposition 3. Idea. The distortion identity is source-agnostic; the rate is bounded by the Gaussian's max-entropy property, with equality iff the source itself is Gaussian. Non-Gaussian sources therefore admit a larger feasible α at the same rate.

Step 1 (distortion identity). For $\hat{X}|X \sim \mathcal{N}(\alpha X, U)$ with $Z \perp X$,

$$D \;=\; (1-\alpha)^2\sigma^2 + U$$

is identity-true (no Gaussianity of X used).

Step 2 (rate via max-entropy). Write $R(\alpha, U; X) = h(\alpha X + Z) - h(Z)$. Since $\mathrm{Var}(\alpha X + Z) = \alpha^2\sigma^2 + U$ for any X with $\mathrm{Var}(X) = \sigma^2$, the Gaussian's max-entropy property gives

$$h(\alpha X + Z) \;\le\; \tfrac{1}{2}\log_2\!\big(2\pi e(\alpha^2\sigma^2 + U)\big),$$

with equality iff $\alpha X + Z$ is Gaussian, i.e. iff X is Gaussian. Subtracting $h(Z) = \frac{1}{2}\log_2(2\pi e U)$ yields the displayed rate bound.

Step 3 (envelope). At fixed R_0, $R(\alpha, U; X)$ is monotone in α at fixed U, so $\alpha_X(R_0, U) \ge \alpha_{\mathrm{Gauss}}(R_0, U)$; since D is strictly decreasing in α on $[0, 1]$, $D^\star_X(R, U) \le D^\star_{\mathrm{Gauss}}(R, U)$. □

B Derivation of the Spatial-URD Functional

Let $\hat{y} = \lfloor g_a(x) \rceil$ and perturb the latent as $\hat{y}_p \mapsto \hat{y}_p + \eta_p$, $\eta_p \sim \mathcal{N}(0, U(p) I_C)$ independently across positions p. With $J_{g_s}[p]$ the decoder Jacobian, J_f the Jacobian of the downstream network f and $H_{\mathcal{L}}$ the Hessian of the task loss $\mathcal{L}$ with respect to the network output, the first-order term vanishes in expectation and, dropping the curvature of f and g_s (Gauss-Newton approximation),

$$\mathbb{E}_\eta[\Delta\mathcal{L}] \;\approx\; \tfrac{1}{2}\sum_p U(p)\, w_f(p), \qquad w_f(p) \;=\; \mathrm{tr}\big(J_{g_s}[p]^\top J_f^\top H_{\mathcal{L}} J_f J_{g_s}[p]\big) \;\ge\; 0,$$

i.e. $\Phi_{w_f}[U] = \int w_f(p) U(p)\, dp$. Its empirical-Fisher proxy $\|\nabla_x \ell\|_p^2$ is very heavy-tailed on natural images (99th/50th percentile ratio between ~ 250 and $\sim 7\times 10^6$ on example images) and concentrates on object boundaries. Φ_{fg} replaces w_f with the indicator of a region; for per-pixel cross-entropy classifiers we take the GT foreground, where predictions are easiest to flip, rather than the background class, which is large and confidently predicted. This choice is motivated by the expansion, not derived from it: the top quantile of w_f is a boundary band, not the foreground. It is validated empirically (Sec. 4.2), including the finding that adding the gradient term on top of the mask does not help (54.7% vs. 54.6%). The noise levels used ($\tau_{\mathrm{base}}/\sqrt{p}$ up to several quantisation steps) are outside the second-order regime, so the expansion explains the form of the predictor, not its accuracy.

C Connections to R-D-P and to entropy models

For the Gaussian family $\hat{X}|X \sim \mathcal{N}(\alpha X, U)$, the marginal is $\hat{X} \sim \mathcal{N}(0, \alpha^2\sigma^2 + U)$. Perfect perception ($P = 0$ under any scale-invariant divergence) requires $\alpha^2\sigma^2 + U = \sigma^2$. Combined with $R = \frac{1}{2}\log\frac{\sigma^2}{U}$, this gives $U_{P=0}(R) = \sigma^2 2^{-2R}$ and

$$D^\star_{P=0}(R) = \big(1 - \sqrt{1 - 2^{-2R}}\big)^2\sigma^2 + \sigma^2 2^{-2R} > \sigma^2 2^{-2R} = D^\star_{\mathrm{RD}}(R),$$

recovering the known distortion penalty for perfect perception [6]. Comparing $U_{P=0}(R) = \sigma^2 2^{-2R}$ with the parametric optimum $U^\star = \sigma^2(2^{2R} - 1)/2^{4R}$ (Cor. 1): at any rate $R > 0$, $U_{P=0}(R) > U^\star$, meaning perfect perception sits at strictly higher uncertainty than the URD-optimal point within the Gaussian test channel family. A unified Rate-Distortion-Uncertainty-Perception (RDUP) objective $\min\, R + \lambda D + \beta U + \gamma P$ subsumes R-D-P ($\beta = 0$) and URD ($\gamma = 0$) as special cases and exposes the full four-way tradeoff space; characterising the achievable (R, D, U, P) region for general sources is a direction for future work.

Coincidence with the Gaussian R-D-P function. In this family, $\alpha^2\sigma^2 = U(2^{2R} - 1)$ at rate R, so $\mathrm{Var}(\hat{X}) = U\,2^{2R}$ and the W_2^2 perception index $P = (\sigma - \sqrt{U}\,2^R)^2$ depends on U alone. With $s = \sqrt{U}\,2^R = \sigma - \sqrt{P}$,

$$D^\star(R, U) \;=\; \sigma^2 + s^2 - 2\sigma s\sqrt{1 - 2^{-2R}},$$

which is the Gaussian R-D-P function under W_2 [37] on the branch $\sqrt{P} \le \sigma\big(1 - \sqrt{1 - 2^{-2R}}\big)$, i.e. $U \in [U^\star, U_{P=0}]$. On that segment URD and R-D-P are reparametrisations of each other; they differ in what they constrain (conditional versus marginal), and the empirical content specific to URD here is the spatial allocation of U.

Entropy modelling vs. predictive uncertainty. Entropy models are employed to estimate the number of bits required to encode latent representations and influence the rate term in the RD objective [4, 25]. They characterise uncertainty about latent symbols from the encoder's perspective, before decoding. Predictive uncertainty U as defined in this work characterises the variability of the reconstructed signal at the decoder, conditioned on the source or compressed representation. While both involve probabilistic descriptions, entropy modelling governs coding efficiency, whereas U reflects ambiguity in the reconstructed output. Proposition 2 and the empirical results in the body show that minimising entropy or rate does not, in general, minimise predictive uncertainty.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The three contribution bullets in Sec. 1 list the propositions, theorems, and empirical numbers that appear in Sec. 3-4; scope qualifiers (linear-m assumption, task-class-specific Φ_{fg}, Gaussian-source converse) are stated alongside the claims, and the Limitations paragraph of Sec. 4 states what remains open.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The *Limitations and scope* paragraph at the end of Sec. 4 lists them: the Gaussian-source restriction of the closed-form URD surface and the right-arm extremality argument; the linear-mean assumption that still binds the lower branch of Th. 1; the restriction of the two-bowl envelope to its two channel classes (entropy-coded quantisers can undercut it); the fact that $U^\star > 0$ is a property of the test-channel family, while practical codecs attain minimum distortion at $U = 0$; the task-class-specificity of Φ_{fg} and its weak or absent effect on (CLIP, DepthAnything-V2); the prompt-matching scope condition that loose SAM2 prompts violate; the codec architecture coverage (Ballé2018 [4] / cheng2020 [9] only); and the fact that the foundation-model probes only show that the axis exists.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical statements (Prop. 1, Prop. 2, Cor. 1, Lem. 1, Th. 1, Th. 2, Th. 3, Cor. 2, Prop. 3) are numbered with assumptions stated inline; the proof of Prop. 1 is in the main text, Lem. 1 is cited [14], and all other proofs are in App. A, with proof sketches in the body. The paper states no conjectures.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Each experiment specifies the codec checkpoint (Ballé2018 [4] / cheng2020 [9] quality levels), downstream model (DeepLabV3-ResNet101, FCN-ResNet50, SAM2.1-Hiera-Tiny, BLIP-base, CLIP-B/16, DepthAnything-V2-Small, Whisper-tiny.en, Qwen2.5-1.5B-Instruct), dataset and split (Pascal VOC 2012, Cityscapes val, Kodak, LibriSpeech `test-clean`, NYU-Depth-v2, GSM8K), noise injection scales, sample counts, and bootstrap procedures (Sec. 4). All models and datasets are publicly available; all experiment scripts, analysis pipelines, per-sample result files, and the two demos (`theory/`, `demo_diffusion/`) are released at `https://github.com/enzotarta/uncertainty-rate-distortion`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Two self-contained demos are included in the code release at `https://github.com/enzotarta/uncertainty-rate-distortion`: `theory/` (URD bowl + the three converse theorems on a Gaussian source under MSE; runs in $\sim$5s on CPU; the full numerical checks, including the Laplacian source, in about a minute) and `demo_diffusion/` (URD axis on a pretrained CIFAR-10 DDPM via init-noise sweep; $\sim$3min on a single GPU). The full set of experiment scripts (codec evaluation, isolation patterns, Φ-aware fine-tune, foundation-model probes) and analysis pipelines are released there, together with the per-sample result files from which the paper's numbers are recomputed (the SAM2 and NYU-depth results are covered by their scripts only). All datasets used (Pascal VOC 2012, Cityscapes, Kodak, LibriSpeech, NYU-Depth-v2, GSM8K) are publicly available; pretrained model checkpoints (Ballé2018 [4], cheng2020 [9], DeepLabV3, FCN, SAM2, BLIP, CLIP, DepthAnything, Whisper, Qwen2.5, U^2-Net) are loaded directly from the standard public registries (CompressAI, torchvision, Hugging Face).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The body Sec. 4 states quality levels, τ grids, sample counts per (image, τ), and sampling-temperature grids. The Φ-aware codec recipe (Sec. 4.2 and the code release) specifies the saliency-weighted MSE loss, α value, training data (VOC train), and quality range. The code release gives the setup of the CDC, cheng2020 [9], SAM2, BLIP, CLIP, DepthAnything, Encodec-Whisper (including the voice-activity threshold), Whisper, and GSM8K experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: 95% percentile bootstrap confidence intervals are reported for the mean mIoU loss (Sec. 4.1, over images), the median Whisper correlation (Sec. 4.3, over (utterance, T) groups) and the CDC curves (Fig. 6, over images); Figs. 3–5 show distributions, medians and accuracies without error bars. Per-image effects are tested with paired Wilcoxon signed-rank tests; ordering accuracy is tested against the relevant chance baseline and reported together with the frequency of the order the predictor implies; correlation medians are tested with sign tests, with reported p-values.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments fit on a single consumer GPU (RTX 3090, 24 GB) or a CPU host. Per-experiment wall-times: codec evaluations and spatial-isolation sweeps < 2 GPU-hours each; the Φ-aware codec fine-tune is ~ 8 GPU-hours per quality level; SAM2/BLIP/CLIP/DepthAnything/Whisper/Qwen probes are inference-only (< 1 GPU-hour each). Total compute for the paper is well below 100 GPU-hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public benchmarks and pretrained models with their original licences and contributes a theoretical/empirical framework for predictive uncertainty in compression and decoding; no human subjects, no private data, no high-risk asset release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The framework is foundational and not tied to a specific deployment. Positive impact: choosing where decoder uncertainty goes in compression and foundation-model decoding can improve downstream task performance at fixed bitrate, and exposes when constructed-uncertainty pipelines are wasting bits on perceptually unimportant regions. Potential negative impact: Φ-aware allocation depends on a saliency proxy that, if biased toward or away from particular content classes (e.g., demographic groups in segmentation), would propagate that bias into rate allocation; we recommend auditing the saliency proxy before deployment in user-facing pipelines.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No high-risk asset is released. The released checkpoints are fine-tunes of the public Ballé2018 [4] and cheng2020 [9] image codecs, plus small decoder-side variance heads, and inherit their low misuse profile (lossy image compression).

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All third-party assets are cited at point of use: Ballé2018 [4] / cheng2020 [9] (CompressAI, Apache-2.0); DeepLabV3, FCN, BLIP, CLIP, U^2-Net (torchvision / Hugging Face, MIT/BSD/Apache); SAM2.1 (Meta, Apache-2.0); DepthAnything-V2 (Apache-2.0); Whisper (MIT); Qwen2.5 (Apache-2.0); Encodec (CC-BY-NC-4.0 model weights, used here only for evaluation, not redistributed); datasets Pascal VOC 2012, Cityscapes (research-only academic licence), Kodak (public PCD release), LibriSpeech (CC-BY-4.0), NYU-Depth-v2 (research licence), GSM8K (MIT). Use of each asset is consistent with its licence terms; we redistribute none of them, apart from per-sample results and 120×120 foreground masks derived from the VOC annotations.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code release includes two self-contained demos (`theory/` for the theory validation, `demo_diffusion/` for the diffusion task), the full experiment scripts and analysis pipelines used to produce every figure in the paper, together with the Φ-aware codec fine-tune recipe (loss definition, training loop, α value, saliency-mask preprocessing). The codec checkpoints (fine-tunes of the public Ballé2018 [4] and cheng2020 [9] checkpoints) and the decoder-side variance heads are released with documentation, linked from the code release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The Qwen2.5-1.5B + GSM8K experiment (Sec. 4.3) uses an off-the-shelf open-weight LLM as one of the foundation-model decoders evaluated; this is a study *of* an LLM rather than a methodological dependence *on* an LLM. As declared at submission, LLMs were used for editing and to help run experiments; no LLM is part of the proposed methods.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.