MAKING LANGUAGE THE LOW-RATE CODE OF VISION: SUCCESSIVE REFINEMENT FROM WORDS TO PIXELS

Anonymous authors
Paper under double-blind review

ABSTRACT

An ordered visual code lets a short prefix stand in for an image, and the training loss decides what that prefix holds. Under pixel reconstruction the first bits go to the highest-variance directions: we construct a source for which a code that is rate-optimal for pixels at every length carries no caption information in its entire first half. A caption is therefore not the low-rate code of its image unless the code is built to make it one. We cast captioning and text-to-image generation as successive refinement under a distortion measure that changes along the code, from caption log-loss in the early stages to pixel distortion in the last. Putting words first then costs nothing among the caption stages exactly when the captions form a summarization chain, and costs the pixels at most the caption uncertainty left in a pixel-optimal reconstruction, which is large for short prefixes and small at full length. We build RUNG, an ordered code whose prefixes are supervised by a summarization ladder of captions and whose full length reconstructs the image; generation extends the code past the prompt and captioning truncates it. With data and capacity held fixed, ordering by language instead of reconstruction raises prefix–caption information 2.9× and GenEval from 0.76 to 0.83, at a cost of 2.8 dB PSNR at a quarter of the code and 0.1 dB at full length. RUNG-XL (3.4B) reaches 0.88 GenEval and 87.6 DPG-Bench in 32 network evaluations, and 0.91 GenEval when errors are redrawn at the rung where they appear; its fine captions entail 94.1% of the attribute tuples in its coarse ones, and it edits images without inversion.

1 INTRODUCTION

Ordered visual codes make coarse-to-fine a property of the representation rather than of the sampler. Nested dropout (Rippel et al., 2014) and its recent descendants (Bachmann et al., 2025; Wen et al., 2025; Miwa et al., 2025) train one-dimensional tokenizers so that every prefix decodes to a valid image. Such a code has to decide what a reader who stops early should learn, and a reconstruction loss decides by default: the first bits go to whatever reduces pixel error most, which in photographs means illumination and low-frequency colour.

The default matters most where text meets images. Unified models for generation and understanding share one token stream (Wang et al., 2026; Chen et al., 2025), diffuse over it (Yang et al., 2025b; Xin et al., 2025), combine next-token prediction with diffusion (Zhou et al., 2025; Deng et al., 2025), or transport text directly to pixels (Liu et al., 2025; He et al., 2025). All of them treat text and images as peer modalities, which overlooks a large difference in rate: a thirty-word caption carries about a hundred nats under a language model (Appendix B), while a one-dimensional tokenizer spends a few thousand on a 256×256 image (2.5k in ours). A text-to-image model must supply the missing information, and these models supply it without structure, as Gaussian noise, sampling temperature or unmasking order; nothing records which parts of the output the prompt fixed. On GenEval (Ghosh et al., 2023), the baselines in Table 2 that score 0.80 or more reach 0.99–1.00 on single objects, yet all but one fall below 0.65 on counting or on position, attributes that short prompts state explicitly.

The obvious fix is to put the caption at the front of an ordered code, so that generation extends the code and captioning truncates it: successive refinement (Equitz & Cover, 1991; Rimoldi, 1994) across modalities, with language as the low-rate code of vision. Under the distortion measure that ordered codes are trained with, this does not happen by itself; Proposition 2 constructs a source for which a

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.

code that is rate-optimal for pixels at every length carries *no* caption information in its first half. On real images a reconstruction-ordered code keeps 0.68 nats of caption information per caption token at a quarter of its length, against 1.97 when the order is anchored in language (Fig. 2a).

Which information goes first is therefore a choice of distortion measure, made stage by stage: we judge the early stages by how well they determine captions and the last by how well it reconstructs pixels. Words-first then loses nothing among the caption stages exactly when each caption summarizes the next finer one (Proposition 1), and costs the pixels an excess rate bounded by the caption uncertainty that a pixel-optimal reconstruction leaves open (Proposition 3), paid at short prefixes and shrinking as fidelity grows. In return the missing information gets a place, the suffix; generation and captioning act on one object, so caption–image agreement is a property of the code rather than of a cycle loss; and an error can be repaired at the rung that introduced it. Figure 1 shows how RUNG learns such a code and uses it to generate, describe, correct and edit. Our contributions are:

- **A distortion schedule for cross-modal coding.** We treat cross-modal generation as successive refinement under a schedule of distortion measures, one optimum per stage, and give the condition under which a caption ladder costs nothing (Proposition 1). The condition is a rule for building training data; violating it costs 5 GenEval points in our ablation.
- **Both prices, bounded and measured.** A reconstruction order can leave language with nothing (Proposition 2). A language order costs the pixels at most the caption uncertainty left in a pixel-optimal reconstruction, and this cost vanishes with fidelity when captions are locally constant (Proposition 3 and Corollary 1). We measure 2.8 dB at a quarter of the code and 2.9% of the rate at full length, with realism unchanged (Fig. 2).
- **RUNG.** A language-anchored ordered code with rate-matched cuts, deterministic descent, stochastic ascent and rung-level self-correction. With data, capacity and generator matched, the order alone is worth 7 GenEval points. RUNG-XL reaches 0.88 GenEval in 32 network evaluations, level with the reward-tuned Lumina-DiMOO at 64, and 0.91 with rung-level correction at $1.2\times$ compute, where best-of-N with the same verifier reaches 0.89 at $4\times$. The same code captions at a chosen granularity and edits without inversion.

2 RELATED WORK

Unified and coarse-to-fine generation. Autoregressive (Wang et al., 2026; Chen et al., 2025), discrete-diffusion (Yang et al., 2025b; Xin et al., 2025; Shi et al., 2026; Li et al., 2026) and hybrid autoregressive–diffusion (Xie et al., 2025a;b; Zhou et al., 2025; Deng et al., 2025; Wu et al., 2026) unified models place text and image tokens in one sequence and let the model infer how they relate (Zhao et al., 2025a). We place them at two depths of one code, so that the relation is part of the representation. Data-to-data couplings transport text directly to images (Liu et al., 2025; He et al., 2025) and close the entropy gap by injecting noise on the text side (Albergo et al., 2025). We keep the gap and give it a location: the part of the code that the caption does not reach. Next-scale (Tian et al., 2024), bitwise (Han et al., 2025), cascaded (Ramesh et al., 2022) and masked (Chang et al., 2023) generators work coarse-to-fine over a spatial or spectral hierarchy on which text acts uniformly; in ours text occupies one interval of a rate hierarchy.

Ordered and language-aware tokenizers. Nested dropout (Rippel et al., 2014) and Matryoshka embeddings (Kusupati et al., 2022) make prefixes usable on their own, and FlexTok (Bachmann et al., 2025), Semanticist (Wen et al., 2025) and One-D-Piece (Miwa et al., 2025) bring the idea to 1D image tokenizers (Yu et al., 2024). In all of these, our closest antecedents, reconstruction sets the order; Semanticist makes this explicit with a PCA-like ordering by explained variance. A second line makes codes semantic: SPAE and LQAE express images in a language model's vocabulary (Yu et al., 2023; Liu et al., 2023); VILA-U, UniTok, TokenFlow and QLIP align codes with text-aligned features (Wu et al., 2025; Ma et al., 2025; Qu et al., 2025; Zhao et al., 2025b); TA-TiTok conditions detokenization on text (Kim et al., 2025). These methods change what tokens mean but leave the order to reconstruction, and none ties a prefix to a caption of matched rate (cf. Table 1, DINOv3).

Rate–distortion with semantic targets. Successive refinement (Equitz & Cover, 1991; Rimoldi, 1994) asks when one nested code is optimal at every rate for a distortion measure. Remote source coding measures distortion on a hidden variable observed through the source (Dobrushin & Tsybakov, 1962; Witsenhausen, 1980) and becomes the information bottleneck under log-loss (Tishby et al.,

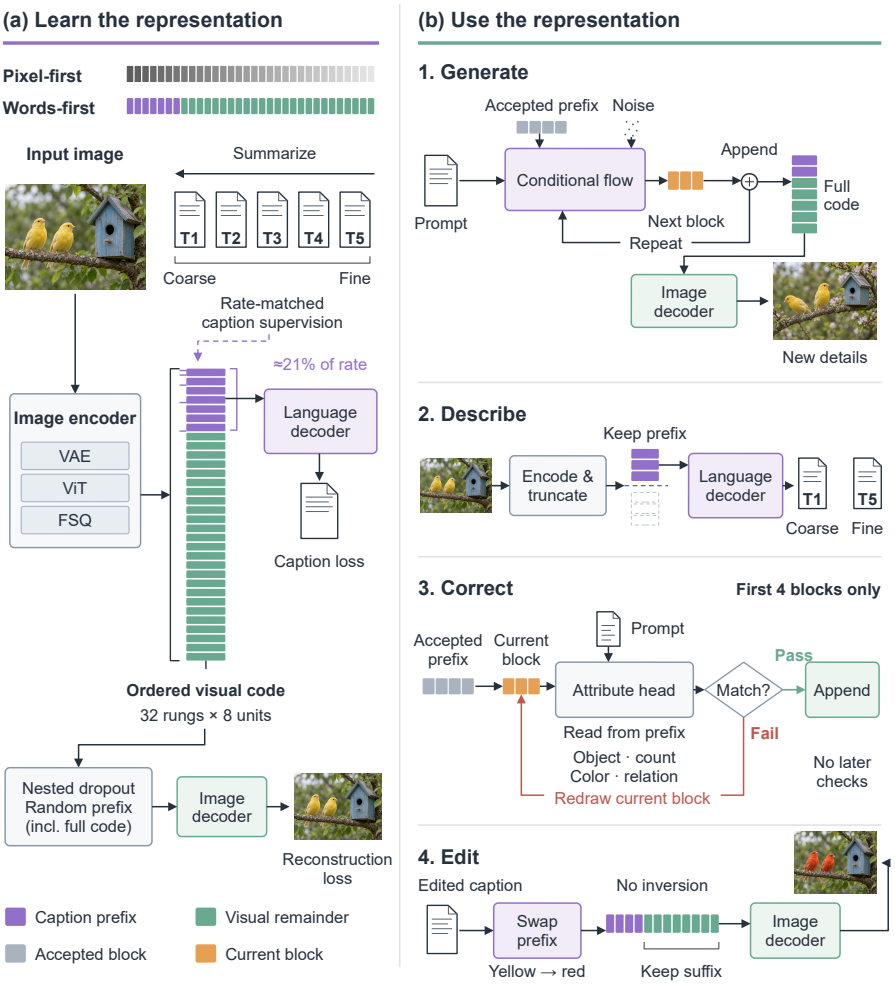

Figure 1: Overview of RUNG. (a) The encoder (VAE, ViT, FSQ) maps an image to an ordered code of 32 rungs of 8 units. Nested dropout trains the image decoder on random prefixes; rate-matched caption supervision trains the language decoder to read each summarized caption level $T_1, \ldots, T_5$ from the prefix whose rate matches it. The captions end within the first $\approx 21\%$ of the rate (purple); the rest holds what they leave free (green). The top bars contrast this words-first order with a reconstruction-ordered, pixel-first code. **(b)** (1) *Generate*: a few-step flow extends the accepted prefix block by block from the prompt and noise (Section 4.2). (2) *Describe*: keep a prefix and decode it, from T_1 to T_5. (3) *Correct*: after each of the first four blocks, an attribute head checks the prefix against the prompt and redraws only a failing block (Section 4.3). (4) *Edit*: swap the caption prefix and keep the suffix, with no inversion (here, recolouring the birds).

1999; Harremoës & Tishby, 2007; Courtade & Weissman, 2014). Semantic rate–distortion constrains semantic and appearance distortion at one rate (Liu et al., 2021); rate–distortion–perception theory adds realism (Blau & Michaeli, 2019; Zhang et al., 2021). Closest to our setting is scalable coding for humans and machines (Choi & Bajić, 2022; Harell et al., 2025), whose base layer serves a machine task. We let the target change *along* one nested code and quantify what each order costs; our base layer is language, which is also the generator's input.

3 Ordering a Code by a Distortion Schedule

3.1 Ordered Codes and Distortion Schedules

Let X be an image (its VAE latent) and $T_1, \ldots, T_M$ captions from coarse to fine. We take captions to be deterministic functions of the image, $T_m = \psi_m(X)$, which holds when a fixed captioner and summarizer decode greedily (Section 5.1); Appendix A.4 treats sampled captions. An ordered

code maps X to $Z = (Z_1, \ldots, Z_L)$, and a reader who stops at k sees the prefix $Z_{\le k}$. Rates are informational, $R_k = I(X; Z_{\le k})$. The successive-refinement coding theorem makes them operational for block codes (Rimoldi, 1994; El Gamal & Kim, 2011), and Section 4.1 measures them for learned codes.

An order is determined by what each stage is judged on. A *distortion schedule* makes this explicit by attaching a distortion measure to every stage and comparing each stage with its own optimum:

- a *caption stage* m is judged by log-loss on T_m. With deterministic captions a prefix is lossless for T_m when it determines T_m, and the least rate at which any code does so is $H(T_m)$;
- a *pixel stage* is judged by a distortion $d(X, \hat{X})$ and compared with the rate–distortion function $R_X(D)$, the least $I(X; \hat{X})$ subject to $\mathbb{E}d(X, \hat{X}) \le D$. The theory uses squared error; Section 5.2 also reports perceptual metrics.

A nested code is *successively refinable for a schedule* if every stage attains its own optimum simultaneously, and the *excess rate* of a stage is its rate beyond that optimum. Classical successive refinement uses one measure throughout; with a pixel measure this is the order of existing ordered tokenizers, which we call *pixel-first*. *Words-first* has caption stages $1, \ldots, M$ followed by pixel stages. Neither order is intrinsic to the image, and Section 3.3 shows that each has a cost.

3.2 When the caption stages cost nothing

Proposition 1 (Caption ladder). *Let a nested code determine T_m from its stage-m prefix for every m. Its stage-m rate is at least $H(T_1, \ldots, T_m)$, an excess of $\Delta_m = H(T_1, \ldots, T_{m-1} \mid T_m)$ over the stage optimum $H(T_m)$, and one code attains this for all m at once. The caption stages are therefore successively refinable if and only if every coarser caption is a function of the next finer one, $T_m = \phi_m(T_{m+1})$. When it is not, stage $m{+}1$ pays at least $\delta_m = H(T_m \mid T_{m+1})$.*

The proof (Appendix A) is the chain rule: a finer prefix contains every coarser one, so it must determine every coarser caption. The proposition is a rule about data. Caption levels should be produced by *summarizing* the next finer level rather than written independently at each length; otherwise the finer stage pays for details that only the coarser caption mentions (Table 8 measures this). The condition is necessary for zero excess rate but does not say what the short levels should contain: truncating one long caption satisfies it exactly and still does worse than summarizing, because the opening words of a long description are not a summary of it.

3.3 The cost of each order

Proposition 2 (Pixel-first can starve language). *For every L there are a source X, a caption $T = \psi(X)$ with $H(T) > 0$, and an ordered code that is successively refinable for X under squared error at each of its L stages, such that $I(Z_{\le k}; T) = 0$ for every $k \le L/2$.*

Let $X = (A, B)$ with A, B independent Gaussians, $\mathrm{Var}(A) \gg \mathrm{Var}(B)$, and $T = \mathbf{1}[B > 0]$. Reverse water-filling spends every nat below $\frac{1}{2}\log(\mathrm{Var}A/\mathrm{Var}B)$ on A, so the first half of an optimal code is independent of T. The construction needs nothing unusual: in photographs the high-variance directions are illumination and low-frequency colour, which captions seldom mention.

Words-first moves the cost to the pixels. Let $T = T_M$; under the condition of Proposition 1 it determines every coarser level. The *price* of putting words first at pixel distortion D is the extra rate needed to determine T and reach D, compared with a code that only reaches D:

$$\pi(D) \;=\; \min_{\mathbb{E}d(X,\hat{X}) \le D} I(X; T, \hat{X}) \;-\; R_X(D). \tag{1}$$

Proposition 3 (Words-first charges pixels, boundedly). *For every D, and any reconstruction $\hat{X}_D^*$ that achieves $R_X(D)$,*

$$\big[H(T) - R_X(D)\big]^+ \;\le\; \pi(D) \;\le\; H\big(T \mid \hat{X}_D^*\big),$$

and $\pi(D) = 0$ if and only if some $\hat{X}_D^*$ determines T. If the pixel stages at $D_1 > \cdots > D_J$ are successively refinable, one words-first code attains the upper bound at every stage simultaneously.

Corollary 1 (The price fades with fidelity). *Suppose a reader $\tilde{\psi}$ and a margin $\rho > 0$ satisfy $\tilde{\psi}(\hat{x}) = \psi(x)$ whenever $d(x, \hat{x}) \le \rho$, for all x outside a set of probability ϵ. Then for $D \le \rho(\frac{1}{2} - \epsilon)$,*

$$\pi(D) \;\le\; h_b\big(\epsilon + D/\rho\big) + \big(\epsilon + D/\rho\big) \log |\mathcal{T}|.$$

If ϵ can be made arbitrarily small by shrinking ρ, that is, if captions are almost surely locally constant, then $\pi(D) \to 0$ as $D \to 0$.

For short prefixes the price is at least $H(T) - R_X(D)$, since any code that determines T spends $H(T)$ nats on it; for long prefixes it is at most what a captioner cannot recover from the best reconstruction, which Corollary 1 drives to zero once reconstructions can be captioned correctly. The corollary is qualitative for long captions, where $\log |\mathcal{T}|$ is large, so Section 5.2 measures the price and tests three predictions. **(P1)** A reconstruction-ordered prefix keeps little caption information. **(P2)** A words-first prefix is less faithful to the pixels at short lengths, and the gap closes with length. **(P3)** The measured price lies between the bounds and is small at full length.

What the order buys. With unlimited capacity every order induces the same $p(x \mid T)$; the order matters for how finite models with short samplers fail. In a words-first code at its stage optimum the prefix up to the prompt's cut k_T is a function of the prompt, $H(Z_{\le k_T} \mid T) = 0$, so an ideal sampler makes no random choice about anything the prompt states, and a partial sample can be read back as a caption and checked against the prompt before later rungs build on it. A learned sampler only approximates this map, which is why the check of Section 4.3 helps. In a pixel-first code the prompt's content is spread over rungs sampled jointly with content the prompt never mentions (tested in Section 5.3).

4 RUNG: A LADDER FROM WORDS TO PIXELS

4.1 BUILDING THE LADDER

Let $h = \mathcal{E}_{\text{vae}}(x)$. RUNG encodes h into an ordered code $z = (z_1, \dots, z_L)$ grouped into L *rungs* and reads it with two decoders: an image decoder that reconstructs h from any prefix, and a language decoder p_ψ that writes a caption from any prefix (Fig. 1a). Generation climbs the rungs and captioning descends them. Training uses, for every image, a caption ladder $T_1, \dots, T_M$ that satisfies Proposition 1 (Section 5.1).

Two terms train the code. The first makes every prefix decodable: the prefix length k is drawn from a nested-dropout distribution q, and the image decoder is trained with flow matching (Lipman et al., 2023) on every truncation,

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{k\sim q}\, \mathbb{E}_{t,\epsilon} \big\| v_\theta(h_t, t, z_{\le k}) - (h - \epsilon) \big\|^2, \qquad h_t = t\, h + (1 - t)\, \epsilon. \tag{2}$$

The second term is what separates RUNG from a reconstruction-ordered tokenizer: each caption level must be recoverable from its matching prefix,

$$\mathcal{L}_{\text{lang}} = \sum_{m=1}^{M} - \log p_\psi\big(T_m \mid z_{\le k_m}\big). \tag{3}$$

Gradients of $\mathcal{L}_{\text{lang}}$ reach the encoder, so language, not variance, decides the competition for early rate. We minimize $\mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{lang}}$. Setting $\lambda = 0$ gives nested dropout, larger λ moves the code toward words-first, and we use $\lambda = 1$ throughout (Fig. 3a in Appendix F traces the trade-off). A third term, introduced in Section 4.2, trains the ascent model.

Rate-matched cuts. Rungs are quantized with FSQ (Mentzer et al., 2024), so the rate of a prefix can be measured. We set $\hat{R}_k = \sum_u \hat{H}(z_u)$, the sum of the empirical entropies of the code units u in the first k rungs; by subadditivity it upper-bounds the prefix entropy. The cut k_m is the first code unit at which $\hat{R}$ reaches $\hat{H}(T_m)$, the caption's entropy under a small language model. This makes the stage optimum of Proposition 1 operational: level m is read where the code has spent the rate that determining T_m requires. The rates depend on the code being trained, so the cuts are recomputed every 5k steps during language anchoring (Appendix C). In the trained code the five levels end at 1.2, 2.9, 6.0, 11 and 21% of $\hat{R}_L \approx 2.5$k nats, inside rungs 1, 1, 2, 4 and 7 of 32 (drawn to scale in Fig. 4, Appendix I).

4.2 DESCENT AND ASCENT

Descent discards information while ascent has to supply it, so the two directions get separate operators instead of one invertible map (Table 9 tests the tied alternative).

Descent (Fig. 1b, *Describe*) is deterministic: $x \mapsto z \mapsto z_{\le k} \mapsto$ greedy decoding of $p_\psi(\cdot \mid z_{\le k})$, one encoder pass and one short decode. Every level reads a prefix of the same code, so the 8-word caption agrees with the 45-word caption for a structural reason; a prompted captioner can only be asked to agree (Table 7).

Ascent (Fig. 1b, *Generate*) is stochastic. Given a prompt T, we locate its cut $k_T = \min\{k : \hat{R}_k \ge \hat{H}(T)\}$ and draw the code in eight blocks. The first four end at the ladder's cuts (rungs 1, 2, 4 and 7) and the last four cover rungs 8–11, 12–16, 17–23 and 24–32. Block b is drawn as $z_{\le k+\Delta_b} \sim p_\theta(\cdot \mid z_{\le k}, T, b)$, conditioned on the prompt and on the blocks before it, by a few-step conditional flow (Frans et al., 2025; Geng et al., 2025) in the rung's pre-quantization space. At the stage optimum the blocks up to k_T are a function of T (Section 3.3), so the flow for these blocks learns a nearly deterministic map and the later blocks carry the randomness. Sampling steps are allocated in proportion to each block's measured rate $\Delta\hat{R}_b$, which is worth 3 GenEval points over a uniform allocation at the same budget (Table 9). A consistency term couples the two directions without requiring invertibility: captions read back from generated prefixes must match the ladder of the training image, $\mathcal{L}_{\text{cons}} = \sum_m -\log p_\psi(T_m \mid \hat{z}_{\le k_m})$.

4.3 CORRECTING ERRORS WHERE THEY APPEAR

Each ladder block can be checked before the next one is drawn (Fig. 1b, *Correct*). We parse the prompt once into attribute tuples (object, count, colour, relation). After each of the first four blocks, an attribute head on the language decoder reads from the prefix the tuples stated by the levels that end in that block; the block is accepted if at least a fraction τ of them agree with the prompt (a tuple the prompt does not mention counts as agreeing) and is otherwise redrawn with fresh noise, up to B times. The check localizes the error. If the first block encodes three apples where the prompt asks for two, the count is caught in that block, because the 8-word level states counts (Appendix B), and only that block is redrawn, where best-of-N would discard the whole sample. Blocks after the last cut are not checked. The verifier is the model's own descent path, with no separate reward model, and never sees the detectors that GenEval uses.

Without training. *Editing* is a prefix swap (Fig. 1b, *Edit*): encode the image, replace the ladder prefix with one drawn from the edited caption, keep the suffix and decode. The prefix carries what the caption states and the suffix what it leaves free, so neither inversion nor attention manipulation is needed; Appendix D lists the cases in which the kept suffix does not fit the new prefix. *Variation at a chosen granularity* is truncation: cut at rung k and resample the rest; the cut decides what is preserved.

5 EXPERIMENTS

5.1 SETUP

Benchmarks. For generation we use GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), T2I-CompBench++ (Huang et al., 2025), TIIF-Bench (Wei et al., 2025), WISE (Niu et al., 2026), and FID and CMMD (Jayasumana et al., 2024) on MJHQ-30K. For understanding we use CapArena-Auto (Cheng et al., 2025), DCScore on DeCapBench (Ye et al., 2025) and CHAIR (Rohrbach et al., 2018). We also report GRANBENCH, 600 prompts whose attributes are annotated with the granularity at which they should appear, and *nesting consistency* (NCC), the fraction of attribute tuples in a model's coarse caption that its own finer caption of the same image entails.

Baselines. We compare with generation-only models (Podell et al., 2024; Esser et al., 2024; Black Forest Labs, 2024; Qin et al., 2025), unified autoregressive and hybrid models (Wang et al., 2026; Chen et al., 2025; Xie et al., 2025b; Deng et al., 2025; Wu et al., 2026) and unified discrete-diffusion models (Yang et al., 2025b; Shi et al., 2026; Xin et al., 2025); with VLMs (Bai et al., 2025; Zhu et al., 2025) as unmatched references for captioning; and with two ordered codes, FlexTok (Bachmann et al., 2025) and Semanticist (Wen et al., 2025), each trained with the same text-to-image head (the

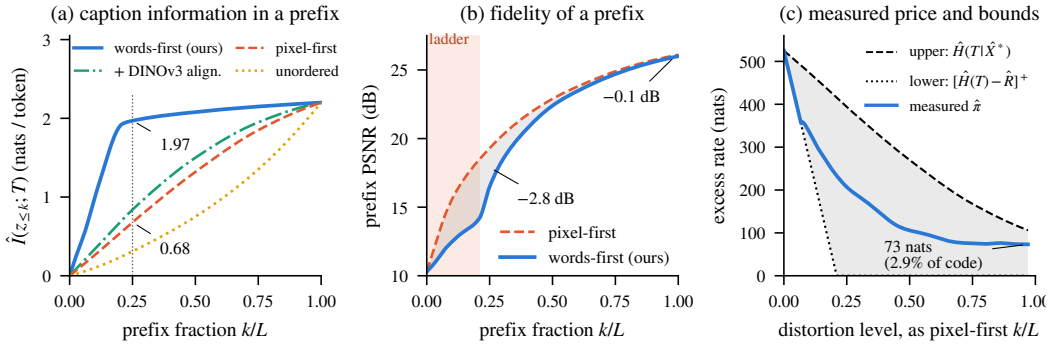

Figure 2: Prefix information, prefix fidelity and the price of words-first. Codes share architecture, data and rate ($\hat{R}_L \approx 2.5$k nats); RUNG-L at full scale. **(a)** Caption information in a prefix, in nats per caption token (P1). **(b)** Pixel fidelity of a prefix: words-first is less faithful inside the ladder (shaded, the first fifth) and then catches up (P2). **(c)** Measured price against the bounds of Proposition 3: on the lower bound while the words-first code is inside its ladder, below the upper bound everywhere, and 73 nats at the right end, where the pixel-first code reaches the words-first code's full-length PSNR (P3).

Table 1: Effect of the ordering signal. RUNG-L at full scale; only the ordering signal changes. $\hat{I}_{25}$: prefix–caption information at $k/L = \frac{1}{4}$ in nats per caption token. Reconstruction metrics (LPIPS: Zhang et al., 2018) use a quarter of the code and all of it; unordered prefixes are not trained to decode. gen. FID: text-to-image on MJHQ-30K.

	caption information and use			quarter of the code			full code		
Ordering signal	$\hat{I}_{25}\uparrow$	GenEval↑	NCC↑	PSNR↑	LPIPS↓	rFID↓	PSNR↑	rFID↓	gen. FID↓
None (unordered)	0.31	0.71	22.1	–	–	–	26.2	0.70	6.8
Nested dropout (pixel-first)	0.68	0.76	41.5	19.3	0.36	2.3	26.1	0.71	6.3
+ DINOv3 alignment	0.84	0.78	48.3	18.8	0.38	2.2	26.0	0.72	6.3
+ language (words-first)	**1.97**	**0.83**	**94.1**	16.5	0.45	2.1	26.0	0.74	6.2

RUNG-L ascent model), data and capacity as RUNG-L. Baseline numbers marked † are transcribed from the original papers, in the setting without prompt rewriting where both are reported; numbers marked ‡ are our evaluations of official checkpoints.

Implementation. The ladder encoder is a ViT-B over SD3 VAE latents that produces $L = 32$ rungs of 8 code units, quantized with FSQ (levels $[8, 8, 8, 5, 5, 5]$, at most 11.1 nats per unit). The image decoder is a 24-layer DiT (Peebles & Xie, 2023) trained with flow matching, and the language decoder is Qwen3-1.7B (Yang et al., 2025a) with LoRA adapters. The ascent model is an MMDiT (Esser et al., 2024) at two sizes, RUNG-L (1.3B) and RUNG-XL (3.4B); the encoder and both decoders are shared, so the two sizes caption identically. Training has three stages: ladder pretraining, language anchoring on LadderCap, and ascent at 256^2 and then 512^2 (Appendix C); RUNG-XL takes about 9 days on 64 A100-80GB GPUs. Sampling uses 32 network evaluations, 24 for ascent and 8 for the image decoder, with guidance 1.5.

LadderCap. We build $M = 5$ caption levels (≈8, 20, 45, 90 and 180 words) for 45M images drawn from Recap-DataComp (Li et al., 2025), BLIP3-KALE (Awadalla et al., 2024), PixelProse (Singla et al., 2024), DOCCI (Onoe et al., 2024), ImageInWords (Garg et al., 2024) and DCI (Urbanek et al., 2024). Following Proposition 1, the longest level is generated from the image (Qwen2.5-VL-7B, greedy) and every shorter level is a *summary of the next longer level* (Qwen3-8B, greedy), never a fresh caption. The summarizer is told to keep a count or a relation in the 8-word level whenever the image contains more than one object. Appendix B reports per-level statistics and a 5K-image entailment audit.

5.2 MEASURING THE ORDER AND ITS PRICE

For Table 1 and Fig. 2 and Fig. 3a we retrain the tokenizer and the RUNG-L ascent model at full scale for each ordering signal, with everything else fixed. $\hat{I}$ is a Barber–Agakov lower bound (Barber &

Agakov, 2003) on the information between a prefix and the finest caption $T = T_5$, in nats per caption token ($\hat{H}(T_5)/|T_5| = 2.5$); an InfoNCE bound (Poole et al., 2019) ranks the codes identically, and the measured price $\hat{\pi}(D)$ and its bounds are defined in Appendix J.

P1: reconstruction order starves language. At a quarter of the code the reconstruction-ordered prefix keeps 0.68 nats of caption information per token and the language-ordered prefix 1.97 (Fig. 2a). Aligning the encoder to DINOv3 features (Siméoni et al., 2025) in the style of REPA (Yu et al., 2025) raises the first number to 0.84: visual self-supervision moves the order toward language, but only a small part of the way. At full length every code determines the image and the four estimates meet at the probe's ceiling of 2.2 nats per token, so the probe measures information, not how readable a particular code is.

P2: words-first costs pixel fidelity where the words are. The words-first prefix is 2.8 dB less faithful at a quarter of the code (Table 1), and the gap is largest inside the ladder, while the code is still encoding captions (Fig. 2b). The cost falls on fidelity to the particular image, not on realism: because the decoder samples, quarter-code decodes remain realistic (rFID 2.1 against 2.3; cf. Blau & Michaeli, 2019). The gap closes to 0.5 dB at half the code and 0.1 dB at full length.

P3: the price stays inside the bounds and is small at full length. Measured in rate (Fig. 2c), the price equals the lower bound $[\hat{H}(T) - \hat{R}]^+$ while the words-first code is inside its ladder, which holds by construction (Appendix J), and stays below the estimated upper bound $\hat{H}(T \mid \hat{X}^*)$ everywhere (207 against 394 nats at a quarter). It ends at 73 nats, 2.9% of the code. Full-length reconstruction (rFID 0.74 against 0.71) and text-to-image FID (6.2 against 6.3) differ by less than the variation across seeds (0.03 and 0.1), and neither application uses the pixels of a short prefix.

The order is a dial. Figure 3a (Appendix F) sweeps λ. Each step toward language buys caption information and GenEval at the expense of quarter-code PSNR. GenEval saturates at λ=1; at λ=3 the code gains little more caption information and starts to lose fidelity at full length (PSNR 25.7). DINOv3 alignment lies inside the trade-off curve: it buys caption information at a worse rate of exchange than supervision with language does. With everything else matched, the order accounts for 7 of the 16 GenEval points between RUNG-L and FlexTok with the same head (0.76 to 0.83 in Table 1); the other 9 come from the rest of the design (Table 9).

5.3 WHAT THE ORDER BUYS

Generation. RUNG-XL ties Lumina-DiMOO at 0.88 GenEval (Table 2) with half its sampling-time parameters (4.0B including the image decoder, against 8B; Appendix C), half its network evaluations (32 against 64) and no reward tuning, and rung-level correction raises it to 0.91. It does *not* lead in every category: on two-object prompts it trails OmniGen2, SD3-Medium, BAGEL and Lumina-DiMOO, and on colour it trails Show-o2, as one would expect from a smaller generator on properties that improve with capacity. Against the models without reward tuning it gains in counting ($+0.04$ over BAGEL), position ($+0.03$ over Janus-Pro) and attribute binding ($+0.03$ over OmniGen2): the three properties a short prompt states explicitly, which the words-first order places in the prompt's prefix (Section 3.3). Lumina-DiMOO is ahead of RUNG-XL on position (0.85 against 0.82) and level on counting; with correction RUNG-XL exceeds every baseline on all three (0.89, 0.87 and 0.84). T2I-CompBench++ (Table 6, Appendix E) shows the same pattern: the gain over Lumina-DiMOO is concentrated in numeracy ($+0.052$) and 2D spatial relations ($+0.048$) and is close to zero for non-spatial relations, which short captions rarely pin down. Extra compute is better spent on rungs than on samples: correction reaches 0.91 at $1.2\times$ the network evaluations and best-of-N with the same verifier 0.89 at $4\times$ (Fig. 3b, Appendix F), since a redraw keeps the blocks that already passed.

Understanding. With a 1.7B language decoder, descent trails the 7–8B VLMs on detail (CapArena-Auto -24.9 against -12.3 and -18.7), lies between them on hallucination, and is ahead of both unified models on every column (Table 7 in Appendix G). On nesting it differs from all prompted models: its fine caption entails 94.1% of the attribute tuples in its coarse caption, against 45–61%, because the two captions read two prefixes of one code instead of coming from two independent decodes. Most of the remaining 5.9% comes from the language decoder (tuples read directly from the prefixes nest 99.2% of the time). On GRANBENCH it scores 71.3, against at most 44.7.

Editing and transfer. A prefix swap edits PIE-Bench images with higher CLIP similarity and lower structure distance than BAGEL's editing mode and RF-Inversion, at one encode and one decode;

Table 2: GenEval. †: transcribed from the original paper, without prompt rewriting (–: not reported). Grey: ordered-code baselines trained by us with the RUNG-L text-to-image head, data and capacity. #Par: generator at sampling time, per each paper's convention (RUNG's totals: Appendix C). *Lumina-DiMOO's last training stage runs GRPO on prompts that follow GenEval's templates.

	Model	#Par	Single	Two	Count	Colors	Pos.	Attr.	Overall
Gen.	SDXL†	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	SD3-Medium†	2B	0.99	0.94	0.72	0.89	0.33	0.60	0.74
	Lumina-Image 2.0†	2.6B	–	0.87	0.67	–	–	0.62	0.73
	FLUX.1-dev†	12B	0.98	0.81	0.74	0.79	0.22	0.45	0.66
AR	Emu3-Gen†	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	Janus-Pro†	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Hybrid	Show-o2†	7B	1.00	0.87	0.58	**0.92**	0.52	0.62	0.76
	OmniGen2†	7B	1.00	**0.95**	0.64	0.88	0.55	0.76	0.80
	BAGEL†	14B	0.99	0.94	0.81	0.88	0.64	0.63	0.82
Disc. diff.	Muddit†	1B	0.98	0.72	0.54	0.82	0.19	0.41	0.61
	MMaDA†	8B	0.99	0.76	0.61	0.84	0.20	0.37	0.63
	Lumina-DiMOO†*	8B	1.00	0.94	0.85	0.89	0.85	0.76	0.88
	Semanticist + T2I head	1.3B	0.98	0.80	0.49	0.85	0.33	0.45	0.65
	FlexTok + T2I head	1.3B	0.98	0.82	0.51	0.86	0.36	0.47	0.67
	RUNG-L	1.3B	0.99	0.89	0.78	0.87	0.75	0.72	0.83
	RUNG-XL	3.4B	0.99	0.93	0.85	0.90	0.82	0.79	0.88
	RUNG-XL + corr.	3.4B	1.00	0.94	**0.89**	**0.92**	**0.87**	**0.84**	**0.91**

on MIMIC-CXR reports, where the impression summarizes the findings, the same two-level ladder raises RadGraph-F1 from 0.312 to 0.331 over a matched baseline (Appendix D).

Ablations use RUNG-L at 256^2 on a 12M-image subset (default GenEval 0.79, against 0.83 at full scale). *The data rule matters as much as the loss* (Table 8, Appendix G). Independently written levels raise $\hat{\delta}$ from 0.12 to 0.83 and cost 5 GenEval points and 31 points of NCC, as Proposition 1 predicts. Truncating one long caption satisfies the chain condition but loses 3 points, because its short levels open a long description without summarizing it (Section 3.2). Design ablations (Appendix H) show saturation at $L = 32$ and a 4-point loss from tying the two directions into one reversible operator.

6 LIMITATIONS AND CONCLUSION

Limitations. A words-first prefix is a worse picture of the particular image than a pixel-first prefix of the same length, so applications that need faithful pixels from few bits should use a small λ or none. The theory assumes deterministic captions and informational rates, Corollary 1 is qualitative for long captions, and our checks rely on estimators that are themselves bounds. LadderCap inherits the biases of the language models that build it, and baseline protocols differ in resolution and prompt handling. Without a human preference study or training at 1024^2, we make no claim about image quality against generation-only systems.

Conclusion. An ordered code always decides what its first bits describe. Left to a reconstruction loss, variance decides and can leave language with nothing (Proposition 2). A distortion schedule makes the choice explicit: for deterministic, locally constant captions, words-first is free when captions summarize one another and reconstructions already determine them, and otherwise costs rate at short prefixes and little at full length. Choosing the schedule per image, where the words end and the pixels begin, is left to future work.

AI USE STATEMENT

Generative AI tools were used only to polish the English of this manuscript. The authors take full responsibility for its content.

Ethics statement

RUNG inherits the risks of its class: representational bias absorbed from web-scale caption data and potential misuse for deceptive imagery. Control over granularity adds no new capability here, although the prefix-swap edit makes targeted semantic edits cheaper, which matters for provenance tooling. The MIMIC-CXR experiment uses a credentialed dataset under its data use agreement and reports a structural transfer result; nothing in it is validated for clinical use.

Reproducibility statement

The ladder objectives are in Section 4.1, ascent and self-correction in Sections 4.2 and 4.3, and complete proofs of Propositions 1 to 3 and Corollary 1, with all assumptions, in Appendix A. The estimators for prefix information, rate and the price bounds are specified in Appendix J. Architecture, staging and hardware are in Section 5.1 and all hyperparameters in Appendix C. LadderCap uses public datasets and public models; prompts, the entailment audit and per-level statistics are in Appendix B. Code, the LadderCap pipeline, GRANBENCH, the NCC evaluator and the scripts that produce Fig. 2 will be released.

References

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Anas Awadalla, Le Xue, Manli Shu, An Yan, et al. BLIP3-KALE: Knowledge augmented large-scale dense captions. *arXiv preprint arXiv:2411.07461*, 2024.

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. FlexTok: Resampling images into 1D token sequences of flexible length. In *International Conference on Machine Learning (ICML)*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

David Barber and Felix Agakov. The IM algorithm: A variational approach to information maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2003.

Black Forest Labs. FLUX.1. https://github.com/black-forest-labs/flux, 2024.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning (ICML)*, 2019.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning (ICML)*, 2023.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. CapArena: Benchmarking and analyzing detailed image captioning in the LLM era. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Hyomin Choi and Ivan V. Bajić. Scalable image coding for humans and machines. *IEEE Transactions on Image Processing*, 31, 2022. doi: 10.1109/TIP.2022.3160602.

Thomas A. Courtade and Tsachy Weissman. Multiterminal source coding under logarithmic loss. *IEEE Transactions on Information Theory*, 60(1):740–761, 2014.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Roland L. Dobrushin and Boris S. Tsybakov. Information transmission with additional noise. *IRE Transactions on Information Theory*, 8(5):293–304, 1962.

Abbas El Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, 2011.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations (ICLR)*, 2025.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. ImageInWords: Unlocking hyper-detailed image descriptions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Alon Harell, Yalda Foroutan, Nilesh Ahuja, Parual Datta, Bhavya Kanzariya, V. Srinivasa Somayazulu, Omesh Tickoo, Anderson de Andrade, and Ivan V. Bajić. Rate-distortion theory in coding for machines and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5501–5519, 2025.

Peter Harremoës and Naftali Tishby. The information bottleneck revisited or how to choose a good distortion measure. In *IEEE International Symposium on Information Theory (ISIT)*, 2007.

Ju He, Qihang Yu, Qihao Liu, and Liang-Chieh Chen. FlowTok: Flowing seamlessly across text and image tokens. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, et al. RadGraph: Extracting clinical entities and relations from radiology reports. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR database (version 2.1.0). PhysioNet, 2024.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. PnP inversion: Boosting diffusion-based editing with 3 lines of code. In *International Conference on Learning Representations (ICLR)*, 2024.

Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Lavida-O: Elastic large masked diffusion models for unified multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Xianhang Li, Haoqin Tu, Mude Hui, et al. What if we recaption billions of web images with LLaMA-3? In *International Conference on Machine Learning (ICML)*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Hao Liu, Wilson Yan, and Pieter Abbeel. Language quantized AutoEncoders: Towards unsupervised text-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jiakun Liu, Wenyi Zhang, and H. Vincent Poor. A rate-distortion framework for characterizing semantic information. In *IEEE International Symposium on Information Theory (ISIT)*, 2021.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. UniTok: A unified tokenizer for visual generation and understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations (ICLR)*, 2024.

Keita Miwa, Kento Sasaki, Hidehisa Arai, Tsubasa Takahashi, and Yu Yamaguchi. One-D-Piece: Image tokenizer meets quality-controllable compression. In *ICML 2025 Workshop on Tokenization (TokShop)*, 2025.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kun-Peng Ning, Bin Zhu, and Li Yuan. WISE: World knowledge-informed semantic evaluation for text-to-image generation. In *International Conference on Machine Learning (ICML)*, 2026.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of connected and contrasting images. In *European Conference on Computer Vision (ECCV)*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning (ICML)*, 2019.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Yu Qiao, Bo Zhang, Xiaohong Liu, Hongsheng Li, Chang Xu, and Peng Gao. Lumina-Image 2.0: A unified and efficient image generative framework. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20031–20042, 2025.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. TokenFlow: Unified image tokenizer for multimodal understanding and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

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

Oren Rippel, Michael A. Gelbart, and Ryan P. Adams. Learning ordered representations with nested dropout. In *International Conference on Machine Learning (ICML)*, 2014.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2025.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng Yan. Beyond text-to-image: Liberating generation with a unified discrete diffusion model. In *International Conference on Learning Representations (ICLR)*, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*, 2024.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

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 (NeurIPS)*, 2024.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control and Computing*, 1999.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating CLIP-style models on dense captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, et al. Multimodal learning with next-token prediction for large multimodal models. *Nature*, 650:327–333, 2026. doi: 10.1038/s41586-025-10041-x.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. TIIF-Bench: How does your T2I model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Xin Wen, Bingchen Zhao, Ismail Elezi, Jiankang Deng, and Xiaojuan Qi. "Principal Components" enable a new language of images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Hans S. Witsenhausen. Indirect rate distortion problems. *IEEE Transactions on Information Theory*, 26(5):518–521, 1980.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ziyi Xia, Ze Liu, Chaofan Li, Haoge Deng, Kun Luo, Bo Zhang, Jiajun Zhang, Dong Liu, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Towards instruction-aligned multimodal generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21964–21975, 2026.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-U: A unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2025a.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-DiMOO: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A. Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, Kevin Murphy, Alexander G. Hauptmann, and Lu Jiang. SPAE: Semantic pyramid AutoEncoder for multimodal generation with frozen LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

George Zhang, Jingjing Qian, Jun Chen, and Ashish Khisti. Universal rate-distortion-perception representations for lossy compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025a.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. QLIP: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *arXiv preprint arXiv:2502.05178*, 2025b.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations (ICLR)*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A PROOFS

Throughout, rates are informational and all random variables take values in standard Borel spaces with the minimizers below attained (true for finite alphabets and for Gaussian sources under squared error). Captions are deterministic, $T_m = \psi_m(X)$, so $H(T_m \mid X) = 0$ and $I(X; T_m) = H(T_m)$. A caption stage is the per-letter measure $d_m(x, \hat{t}) = \mathbf{1}[\hat{t} \neq \psi_m(x)]$ at distortion level zero, whose optimum is $H(T_m)$. A nested code is represented by its stage descriptions $W_1, W_2, \ldots$ with W_j a function of W_{j+1}; the successive-refinement coding theorem (Rimoldi, 1994; El Gamal & Kim, 2011) makes the cumulative rates $I(X; W_j)$ achievable by one block code. Its proof (superposition coding with a joint-typicality check per stage for achievability, and the chain rule for the converse) uses the distortion constraint of each stage separately and never uses that the stages share a measure, so it applies to a schedule whose stages are judged by different measures.

A.1 PROOF OF PROPOSITION 1

Converse. Let W_m be the stage-m description, which determines T_m. Since W_j is a function of W_m for $j \leq m$, W_m determines $(T_1, \ldots, T_m)$, so $X \to W_m \to (T_1, \ldots, T_m)$ is a Markov chain. By data processing, $I(X; W_m) \geq I(X; T_1, \ldots, T_m) = H(T_1, \ldots, T_m)$. The excess over the stage optimum $H(T_m)$ is $H(T_1, \ldots, T_m) - H(T_m) = H(T_1, \ldots, T_{m-1} \mid T_m) = \Delta_m$.

Achievability. Take $W_m = (T_1, \ldots, T_m)$. These descriptions are nested, determine T_m, and satisfy $I(X; W_m) = H(T_1, \ldots, T_m)$ for every m simultaneously.

Characterization. $\Delta_m = 0$ for all m iff each T_j, $j < m$, is a function of T_m for all m, which holds iff $T_m = \phi_m(T_{m+1})$ for every m (one direction by composition, the other by taking $j = m - 1$). Finally $\Delta_{m+1} = H(T_1, \ldots, T_m \mid T_{m+1}) \geq H(T_m \mid T_{m+1}) = \delta_m$. □

A.2 PROOF OF PROPOSITION 2

Let $A \sim \mathcal{N}(0, \sigma_A^2)$ and $B \sim \mathcal{N}(0, \sigma_B^2)$ be independent, $X = (A, B)$, $d(x, \hat{x}) = \|x - \hat{x}\|^2$, and $T = \mathbf{1}[B > 0]$, so $H(T) = \log 2$. For a Gaussian vector under squared error, reverse water-filling (Cover & Thomas, 2006, Ch. 10) gives the optimal test channel at rate R: with water level θ, component A receives rate $\frac{1}{2}\log(\sigma_A^2/\theta)$ and B receives $\max\{0, \frac{1}{2}\log(\sigma_B^2/\theta)\}$. For $R \leq R^\star = \frac{1}{2}\log(\sigma_A^2/\sigma_B^2)$ the water level satisfies $\theta \geq \sigma_B^2$, so B receives no rate and the optimal reconstruction is $(\hat{A}, 0)$ with $\hat{A}$ a function of A and of noise independent of B. Gaussian sources are successively refinable under squared error (Equitz & Cover, 1991), and the reverse water-filling solutions at decreasing distortions are nested, so a single code attains $R_X(D_j)$ at every stage. Let the stages have rates $r, 2r, \ldots, Lr$ and choose $\sigma_A^2/\sigma_B^2 \geq e^{Lr}$, so that $R^\star \geq Lr/2$. Every stage $k \leq L/2$ then has rate at most $R^\star$, its description is independent of B, and $I(Z_{\leq k}; T) \leq I(Z_{\leq k}; B) = 0$. □

The argument uses only that the caption depends on low-variance directions; any source whose caption-relevant directions carry less variance than caption-irrelevant ones behaves the same way at low rate.

A.3 PROOF OF PROPOSITION 3 AND COROLLARY 1

Any description W from which T is determined and $\hat{X}$ is reconstructed satisfies $I(X; W) \geq I(X; T, \hat{X})$ by data processing, and $(T, \hat{X})$ is itself such a description; hence $R_X(D) + \pi(D)$ in equation 1 is the least rate of a code that determines T and reaches distortion D.

Lower bounds. For any feasible $\hat{X}$, $I(X; T, \hat{X}) \geq I(X; \hat{X}) \geq R_X(D)$, so $\pi(D) \geq 0$; and $I(X; T, \hat{X}) \geq I(X; T) = H(T)$, so $\pi(D) \geq H(T) - R_X(D)$.

Upper bound. For an optimal $\hat{X}_D^*$, $I(X; T, \hat{X}_D^*) = I(X; \hat{X}_D^*) + I(X; T \mid \hat{X}_D^*) = R_X(D) + H(T \mid \hat{X}_D^*) - H(T \mid X, \hat{X}_D^*) = R_X(D) + H(T \mid \hat{X}_D^*)$, since T is a function of X.

Equality case. If some optimal $\hat{X}_D^*$ determines T, the upper bound is 0. Conversely, if $\pi(D) = 0$, let $\hat{X}^\circ$ attain the minimum in equation 1. Then $R_X(D) = I(X; T, \hat{X}^\circ) = I(X; \hat{X}^\circ) + H(T \mid$

$\hat{X}^\circ) \geq R_X(D) + H(T \mid \hat{X}^\circ)$, so $H(T \mid \hat{X}^\circ) = 0$ and $I(X; \hat{X}^\circ) = R_X(D)$: $\hat{X}^\circ$ is an optimal reconstruction that determines T.

Simultaneity. Let $X \to \hat{X}_J \to \cdots \to \hat{X}_1$ be a successively refinable pixel code with $I(X; \hat{X}_j) = R_X(D_j)$ and $\mathbb{E}d(X, \hat{X}_j) \leq D_j$. Define the words-first descriptions $W_0 = T$ and $W_j = (T, \hat{X}_1, \ldots, \hat{X}_j)$. They are nested and $I(X; W_j) = I(X; \hat{X}_{1:j}) + H(T \mid \hat{X}_{1:j}) = I(X; \hat{X}_j) + H(T \mid \hat{X}_{1:j}) \leq R_X(D_j) + H(T \mid \hat{X}_j)$, using the Markov chain for the first term and the fact that conditioning reduces entropy for the second. Preceding W_0 by the caption stages of Proposition 1 (zero excess under the chain condition) gives one words-first code that meets every stage's upper bound. □

Corollary. Let $P_e = \Pr[\tilde{\psi}(\hat{X}^*_D) \neq T]$. An error requires either that X lie in the exceptional set or that $d(X, \hat{X}^*_D) > \rho$, so by the union bound and Markov's inequality $P_e \leq \epsilon + \mathbb{E}d(X, \hat{X}^*_D)/\rho \leq \epsilon + D/\rho \leq \frac{1}{2}$. Conditioning on $\hat{X}^*_D$ reveals at least as much as conditioning on $\tilde{\psi}(\hat{X}^*_D)$, so by Fano's inequality $H(T \mid \hat{X}^*_D) \leq h_b(P_e) + P_e \log(|\mathcal{T}| - 1)$. Since h_b increases on $[0, \frac{1}{2}]$, $H(T \mid \hat{X}^*_D) \leq h_b(\epsilon + D/\rho) + (\epsilon + D/\rho) \log |\mathcal{T}|$, and Proposition 3 gives the bound on $\pi(D)$. For the limit, fix $\epsilon > 0$ and a matching $\rho(\epsilon)$; then $\limsup_{D \to 0} \pi(D) \leq h_b(\epsilon) + \epsilon \log |\mathcal{T}|$, which tends to zero with ϵ because $\mathcal{T}$ is finite. □

A.4 Sampled captions

If the captioner samples, $T_M \sim P(\cdot \mid X)$ and $T_m = \phi_m(T_{m+1})$, a prefix can at best be *sufficient* for T_m, $I(Z_{\leq k_m}; T_m) = I(X; T_m)$. The stage optimum becomes the rate of the minimal sufficient statistic of X for T_m, and the same arguments go through with $I(X; \cdot)$ in place of $H(\cdot)$; the summarization chain becomes the Markov chain $X \to T_M \to \cdots \to T_1$. LadderCap decodes greedily, so the deterministic case of the main text is the one that applies.

B LadderCap construction and audit

Prompts. The base captioner (Qwen2.5-VL-7B, greedy, at most 256 new tokens) receives *"Describe this image in about 180 words. Mention every visible object with its count, colour and position, then the setting, lighting and style."* The summarizer (Qwen3-8B, greedy, thinking disabled) receives, for each level $m < M$, the level-$(m{+}1)$ caption and *"Summarize the description in at most N_m words. Use only facts stated in it. Keep counts, colours and spatial relations of the main objects before any other detail."* with $N_m \in \{8, 20, 45, 90\}$.

Level 1 states a count or a relation. When the level-5 caption mentions two or more objects, the 8-word level must state a count or a relation between them. The summarizer is re-prompted once when a rule-based parser finds neither; afterwards 91% of level-1 captions of multi-object images satisfy the rule, and the rest keep the first output.

Statistics. Table 3 lists per-level lengths, caption entropies under the Qwen3-0.6B model of Appendix J, the resulting cuts, and the entailment audit. A thirty-word caption, between levels 2 and 3, carries about a hundred nats.

Entailment audit. On 5K held-out images we extract attribute tuples from every level and ask Qwen2.5-72B-Instruct whether each tuple of level m is entailed by level $m + 1$; two annotators checked 500 images and agreed with the judge on 94% of tuples. Violations are mostly paraphrased colours ("maroon" summarized as "red") and merged counts ("two cups and a mug" summarized as "three cups").

Stability across summarizers. Replacing Qwen3-8B by Llama-3.1-8B-Instruct or Gemma-2-9B-it changes $\hat{\delta}$ from 0.12 to 0.14 and 0.13, GenEval in the ablation setting from 0.79 to 0.79 and 0.78, and NCC from 92.7 to 92.1 and 92.4.

Table 3: LadderCap levels. Tokens use the Qwen3 tokenizer. $\hat{H}(T_m)$: cross-entropy under Qwen3-0.6B fine-tuned on LadderCap (an upper bound on the entropy). Cut: where the trained RUNG code reaches $\hat{H}(T_m)$. Entailed: fraction of level-m attribute tuples entailed by level $m{+}1$ on 5K audited images.

Level	words	tokens	$\hat{H}(T_m)$ (nats)	nats/token	cut (% of $\hat{R}_L$)	cut rung	entailed (%)
T_1	8.1	10	30	3.0	1.2	1	97.9
T_2	19.6	25	72	2.9	2.9	1	97.2
T_3	44.8	57	150	2.6	6.0	2	96.6
T_4	89.5	110	275	2.5	11.0	4	95.8
T_5	178.3	210	525	2.5	21.0	7	–

C HYPERPARAMETERS

Networks. Table 4 lists every network, its size and whether it runs at sampling time. The tables of the main text give the ascent model only (#Par), following the baselines' convention of counting the generator. Counting the image decoder as well, RUNG-L uses 1.9B parameters at sampling time and RUNG-XL 4.0B; with correction the language decoder adds 1.7B (3.6B and 5.7B).

Training. All stages use AdamW ($\beta = (0.9, 0.99)$, weight decay 0.01), bf16, EMA 0.9999 and 64 A100-80GB GPUs. *Ladder pretraining* ($\lambda = 0$): 300k steps, batch 1024, 256^2, learning rate 10^{-4}; the prefix length is $k = L$ with probability 0.2 and otherwise uniform over code units. *Language anchoring* ($\lambda = 1$): 150k steps, batch 512, learning rate $5 \cdot 10^{-5}$ for the code and $2 \cdot 10^{-4}$ for the LoRA adapters (rank 64); cuts are recomputed every 5k steps from running entropy estimates on 100k held-out images. *Ascent*: 300k steps at 256^2 (batch 1024) and 100k steps at 512^2 (batch 512), learning rate 10^{-4}, text dropout 0.1 for guidance; $\mathcal{L}_{\text{cons}}$ is added with weight 0.1 for the last 50k steps, using one-step flow estimates of the generated prefix.

Sampling. Ascent uses 24 flow steps in total, allocated in proportion to block rate with at least one step per block: $(1, 1, 2, 2, 3, 4, 5, 6)$ for the eight blocks of Section 4.2. Classifier-free guidance (1.5) is applied to ascent only, with both branches in one batch. The image decoder uses 8 Euler steps without guidance. This gives 32 network evaluations; VAE decoding is not counted, for RUNG or for any baseline.

Correction. Threshold $\tau = 0.8$ and redraw budget $B = 2$ in all tables. The verifier runs once after each of the four ladder blocks (4 evaluations). On GenEval a sample needs 0.9 redraws on average, each costing 1.5 ascent steps plus one verifier pass, so a corrected sample costs $32{+}4{+}0.9{\times}2.5 \approx 38$ evaluations (1.2×). Figure 3b varies $B \in \{1, 2, 3, 4\}$ with $\tau = 0.8$ for $B \leq 2$ and $\tau = 0.9$ above. Best-of-N draws N full samples and keeps the one whose final caption agrees best with the prompt under the same attribute head.

Table 4: Networks. Sampling: whether the network runs when generating an image.

Network	Architecture	Parameters	Sampling
VAE	SD3 VAE, frozen	84M	decode only
Ladder encoder	ViT-B over VAE latents, FSQ $[8, 8, 8, 5, 5, 5]$	86M	no
Image decoder	DiT, 24 layers, width 1152	0.58B	8 steps
Language decoder	Qwen3-1.7B + LoRA (rank 64) + attribute head	1.7B	correction only
Ascent, RUNG-L	MMDiT, 24 layers, width 1536	1.3B	24 steps
Ascent, RUNG-XL	MMDiT, 32 layers, width 2048	3.4B	24 steps

D TRUNCATION, VARIATION AND EDITING

Table 5 quantifies variation by truncation on 1K held-out images: the code is cut at rung k, the rest is resampled four times, and we measure what survives. Up to rung 7 the preserved attribute tuples follow the caption levels that the prefix contains; past the ladder, later cuts preserve structure and texture while the tuples are already fixed. At rung 20 most of the remaining change is global illumination (mean luminance changes by 9% against 23% at rung 8).

Editing and variation. On PIE-Bench (Ju et al., 2024), a prefix swap reaches CLIP similarity 26.4 (whole image) and 23.1 (edited region) at structure distance 19.6×10^{-3}, against 25.8/22.7 at 27.4 for BAGEL's editing mode and 25.2/22.2 at 40.6 for RF-Inversion (Rout et al., 2025), both run by us. The swap needs one encode and one decode; inversion costs 2–3× more. Cutting at rung 8, one rung past the finest caption, keeps 96% of the 180-word description's attribute tuples; cutting at rung 20 mainly changes lighting (Table 5).

Transfer. A radiology report's impression approximately summarizes its findings, which gives a two-level ladder. On MIMIC-CXR (Johnson et al., 2019; 2024), RUNG reaches 0.331 RadGraph-F1 (Jain et al., 2021) and 0.495 CheXbert-F1 (Smit et al., 2020) with 0.91 two-level consistency, against 0.312/0.478 for a matched baseline that writes the two sections independently (structure only, not clinical accuracy).

Failure cases. We see four recurring failures. (i) *Rendered text.* Captions rarely spell out text, so glyphs fall in the suffix and are resampled; words on signs change under variation and are garbled in generation. (ii) *Large layout edits.* When a prefix swap moves the main object, the kept suffix still encodes texture for the old layout and seams appear; this affects 7% of PIE-Bench edits, mostly in the "change position" category. (iii) *Counts above five.* The 8-word level often writes "several", so correction cannot catch a wrong count of seven. (iv) *Small objects that only the 180-word level names.* They are fixed late in the ladder, where a redraw is cheap but the attribute head is least accurate (82% tuple accuracy at level 5 against 97% at level 1).

Table 5: Truncation and resampling. Tuples kept: fraction of the level-5 attribute tuples that hold in the resampled image (judged as in Appendix B). DINO: structure similarity to the source; LPIPS: distance to the source. Mean over 1K images and four samples.

Cut rung	rate (nats)	levels contained	tuples kept (%)	DINO ↑	LPIPS ↓
2	156	T_1–T_3	41	0.52	0.63
4	312	T_1–T_4	63	0.61	0.56
8	625	T_1–T_5	96	0.74	0.47
20	1560	T_1–T_5	98	0.91	0.24

E INSTRUCTION FOLLOWING AND EFFICIENCY

Table 6 complements Table 2 with DPG-Bench, T2I-CompBench++, TIIF-Bench, WISE, FID and sampling cost.

Table 6: Instruction following, composition and efficiency. GenEval, DPG and WISE baselines are transcribed from the original papers (WISE entries marked ‡ are ours); T2I-C++ (T2I-CompBench++ mean), TIIF, FID (MJHQ-30K), NFE and latency are measured by us with official checkpoints and default samplers. NFE counts forward passes of every sampling network, image decoder and verifier included; guidance branches are batched and counted once, and VAE decoding is excluded for all models. Latency at 512^2, batch 1, one A100.

Model	GenEval	DPG	T2I-C++	TIIF	WISE	FID↓	NFE↓	Lat. (s)↓
SDXL	0.55	74.7	0.408	51.2	0.43	8.8	50	3.1
FLUX.1-dev	0.66	83.8	0.502	72.6	0.50	6.1	50	4.8
Lumina-Image 2.0	0.73	87.2	0.487	69.1	0.43‡	6.8	40	3.4
Janus-Pro	0.80	84.2	0.471	66.3	0.35	8.0	576	9.1
BAGEL	0.82	85.1	0.524	71.4	0.52	6.4	50	7.2
Lumina-DiMOO	0.88	86.0	0.551	74.8	0.60‡	6.9	64	6.3
FlexTok + T2I head	0.67	80.3	0.421	58.7	0.34	7.9	32	1.9
RUNG-L	0.83	85.3	0.538	72.9	0.54	6.2	32	**1.4**
RUNG-XL	0.88	87.6	0.581	76.2	0.61	5.8	32	2.1
RUNG-XL + corr.	**0.91**	**88.4**	**0.607**	**78.5**	**0.63**	**5.6**	38	2.5

F THE ORDERING DIAL AND TEST-TIME COMPUTE

Figure 3 supports two results of Sections 5.2 and 5.3: the trade-off traced by the language weight λ, and the comparison of rung-level correction with best-of-N at matched verifier.

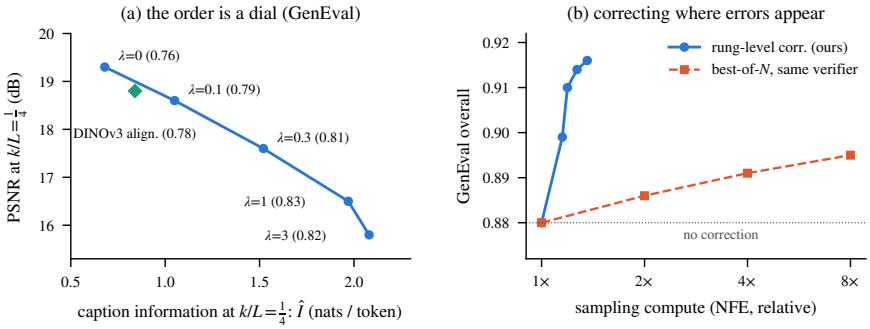

Figure 3: Left: the order is a dial. Sweeping λ trades quarter-code PSNR for caption information (GenEval in parentheses); λ=0 is nested dropout, and DINOv3 alignment lies inside the trade-off curve. **Right: test-time compute.** Rung-level correction (redraw budgets $B = 1$ to 4; Appendix C) against best-of-N with the same verifier, in network evaluations relative to one uncorrected sample.

G CAPTIONING AND LADDER-CONSTRUCTION RESULTS

Table 7 gives the captioning comparison discussed in Section 5.3, and Table 8 the ladder-construction ablation of Proposition 1, in the ablation setting of Section 5.3.

Table 7: Captioning. CapA.: CapArena-Auto (−100 to 100); DCS: DCScore; NCC: nesting consistency of a model's own 8- and 45-word captions (baselines prompted for both); Gran.: GRANBENCH. VLMs are unmatched references; all baseline numbers are ours (‡).

Model	CapA.	DCS	CHAIR$_s$↓	NCC	Gran.
Qwen2.5-VL-7B‡	−12.3	42.8	11.2	61.4	44.7
InternVL3-8B‡	−18.7	39.5	13.5	57.9	41.2
BAGEL‡	−46.5	31.2	18.4	48.2	33.6
Lumina-DiMOO‡	−53.2	29.8	21.0	45.1	31.9
RUNG descent	−24.9	38.6	12.8	**94.1**	**71.3**

Table 8: Building the ladder (Proposition 1), ablation setting. $\hat{\delta}$: estimated $H(T_m \mid T_{m+1})$ in nats per token, averaged over levels. Truncation satisfies the chain condition exactly; its $\hat{\delta}$ is estimator bias.

Construction	$\hat{\delta}$	GenEval	NCC
Independent per length	0.83	0.74	61.8
Truncated long caption	**0.04**	0.76	90.4
Successive summary	0.12	**0.79**	**92.7**

H DESIGN ABLATIONS

RUNG-L at 256^2 on the 12M-image subset, as in Section 5.3. Replacing the learned language decoder with frozen embeddings, the natural choice for an exactly invertible text interface, costs 9 points for a context-free lookup and 12 for a pooled sentence embedding: such embeddings carry no compositional structure, and the ascent model has to recover relations on its own. Tying descent and ascent into one time-reversible operator costs 4 points and 6 points of NCC, which supports treating the asymmetry between the two directions as real rather than as something to engineer away.

Table 9: Design. Shaded rows are the defaults used everywhere else. Latency per image at 256^2.

Variant	GenEval	NCC	Lat. (s)
Ascent parameterization			
Autoregressive over rungs	0.76	–	2.6
One diffusion, full code	0.77	–	7.9
Per-rung few-step flow	**0.79**	–	**0.9**
Step allocation across rungs			
Uniform	0.76	–	0.9
$\propto \Delta\hat{R}_b$	**0.79**	–	0.9
Representation			
$L = 8$ / 16 / **32** / 64	0.75 / 0.78 / **0.79** / 0.79	–	–
Continuous rungs (no FSQ)	0.77	91.0	–
Text interface			
Frozen context-free lookup	0.70	77.6	–
Frozen pooled sentence embedding	0.67	73.1	–
Learned language decoder	**0.79**	**92.7**	–
Coupling of the two directions			
Symmetric, tied / reversed	0.75	86.7	–
Asymmetric descent / ascent	**0.79**	**92.7**	–

I CODE LAYOUT AT MEASURED RATE

Figure 4 draws the two orders of Fig. 1a to scale, with the caption cuts of Section 4.1 placed at their measured rates.

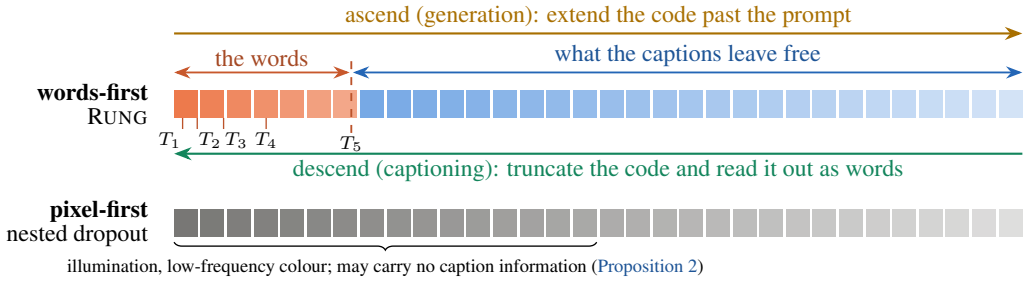

Figure 4: Who gets the first bits, at measured rate. Two ordered codes of equal length and rate, read left to right; shading indicates information per rung. *Top:* RUNG judges its early stages by captions and its last stage by pixels; ticks mark where caption levels $T_1, \ldots, T_5$ end, placed at measured rate (Section 4.1), so the captions occupy the first fifth of the code. *Bottom:* trained by reconstruction alone, the code spends its early rate on high-variance content. Words-first costs the pixels most at short prefixes and about 3% of the rate at full length (Fig. 2).

J ESTIMATORS

Prefix–caption information. $\hat{I}(z_{\le k}; T)$ is the Barber–Agakov bound $\hat{H}(T) - \mathbb{E}[-\log q(T \mid z_{\le k})]$ (Barber & Agakov, 2003) with $T = T_5$ and q a probe decoder (Qwen3-1.7B with LoRA) trained on frozen prefixes of each code. We divide by the mean caption length (210 tokens) and report nats per token. One probe is trained per code and prefix length on 1M held-out images, with the same architecture and budget for all codes. The bound lies below $\hat{H}(T_5)/|T_5| = 2.5$ nats per token by the probe's cross-entropy, which is 0.3 nats per token at full length and gives the ceiling of 2.2 in Fig. 2a. As a check, the InfoNCE bound (Poole et al., 2019) with critic $f(z_{\le k}, T) = \langle g(z_{\le k}), u(T)\rangle/\tau$, where g is a 4-layer transformer over the prefix and u a frozen sentence encoder with a trained linear head, ranks the codes identically at every k; at a quarter of the code it gives 1.1, 2.3, 2.7 and 5.6 nats per caption for the unordered, pixel-first, DINOv3-aligned and words-first codes. With batch 4096 this bound cannot exceed $\log 4096 \approx 8.3$ nats per caption, far below $\hat{H}(T_5) = 525$, so we use it only for ranking.

Rate. $\hat{R}_k = \sum_u \hat{H}(z_u)$ over the code units u in the first k rungs, with $\hat{H}(z_u)$ the plug-in entropy of unit u's FSQ symbol (64,000 values) on 1M images. By subadditivity the sum upper-bounds the joint prefix entropy, which equals $I(X; Z_{\le k})$ for a deterministic encoder. Caption entropies $\hat{H}(T_m)$ are cross-entropies under Qwen3-0.6B fine-tuned on LadderCap, which also upper-bound the true entropies. The two biases move the cuts in opposite directions: over-estimating $\hat{R}$ places a cut too early and over-estimating $\hat{H}(T_m)$ places it too late. Moving every cut by ± 4 code units changes GenEval by at most 0.01 in the ablation setting.

Price. For a target PSNR D, $\hat{R}_{\mathrm{pix}}(D)$ and $\hat{R}_{\mathrm{words}}(D)$ are the prefix rates at which each code first reaches D on 5K held-out images, interpolated between code units. The measured price is $\hat{\pi}(D) = \max\{\hat{R}_{k_M}, \hat{R}_{\mathrm{words}}(D)\} - \hat{R}_{\mathrm{pix}}(D)$, which counts the words-first code as having determined T once it passes its last cut. It is defined up to the words-first code's full-length PSNR (26.0 dB), which the pixel-first code reaches at $k/L = 0.97$; this is the right end of Fig. 2c. The upper bound $\hat{H}(T \mid \hat{X}^*_D)$ is the cross-entropy of a captioning probe (Qwen2.5-VL-3B with LoRA) fine-tuned to predict T from pixel-first reconstructions at the matching prefix. Cross-entropy over-estimates conditional entropy, so a measured price below the estimated bound is a necessary check, not a sufficient one. Consistent with data processing, this image probe recovers slightly less than the prefix probe (0.62 against 0.68 nats per token at a quarter of the code). The lower bound is $[\hat{H}(T) - \hat{R}_{\mathrm{pix}}(D)]^+$. While the words-first code is still inside its ladder, $\hat{R}_{\mathrm{words}}(D) < \hat{R}_{k_M} \approx \hat{H}(T)$ and $\hat{\pi}$ equals this bound by construction, so the informative comparison is with the upper bound.