AUTOREGRESSIVE VISUAL GENERATION NEEDS A PROLOGUE

Anonymous authors
Paper under double-blind review

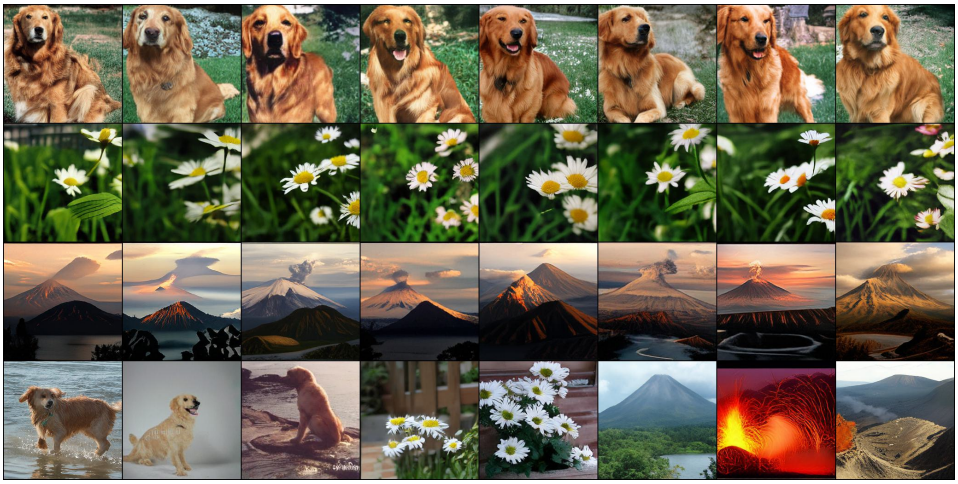


Figure 1: **Top three rows**: each row generated with **fixed prologue tokens** while resampling visual tokens. Samples within a row share remarkably consistent global structure, including viewpoint, spatial composition, color palette, and foreground/background style. This structure **emerges naturally** under the AR compression objective alone, *without* any semantic supervision. **Bottom row**: samples drawn from *different* prologue tokens, shown for reference.

ABSTRACT

In this work, we propose **Prologue**, an approach to bridging the reconstruction–generation gap in autoregressive (AR) image generation. Instead of modifying visual tokens to satisfy both reconstruction and generation, Prologue learns a small set of image-dependent *prologue tokens* prepended to the visual token sequence. These prologue tokens are trained *exclusively* with the AR cross-entropy (CE) loss, while visual tokens remain dedicated to reconstruction. This decoupled design lets us optimize generation through the AR model's true distribution while largely preserving reconstruction quality, which we further formalize from an ELBO perspective. On ImageNet 256×256, Prologue-Base reduces gFID from 21.01 to 10.75 without classifier-free guidance while keeping reconstruction almost unchanged; Prologue-Large reaches a competitive rFID of 0.99 and gFID of 1.46 using a standard AR model without auxiliary semantic supervision. Interestingly, driven only by AR gradients, prologue tokens exhibit emergent semantic structure: linear probing on 16 prologue tokens reaches 35.88% Top-1, far above the 23.71% of the first 16 tokens from a standard tokenizer; resampling with fixed prologue tokens preserves a similar high-level semantic layout. Our results suggest a new direction: generation quality can be improved by introducing a separate learned generative representation while keeping visual tokens dedicated to reconstruction.

1 INTRODUCTION

Tokens learned for reconstruction may still be difficult for an autoregressive (AR) model to predict. This reconstruction–generation gap is a central challenge in discrete image generation. From 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.

theoretical perspective, two-stage discrete image generation (Esser et al., 2021) can be formulated within the evidence lower bound (ELBO) framework (Kingma & Welling, 2014), which includes a reconstruction term and a prior matching term. Under deterministic encoding, the latter reduces to a negative log-prior term, whose dataset average is the cross-entropy between the aggregate posterior and the model prior. This simplifies the training pipeline but introduces the well-known reconstruction-generation gap: the tokenizer never accounts for the learnability of the AR, and the data-model gap is reduced unilaterally by the AR.

Existing work explores several orthogonal directions to improve the AR-compatibility of tokens: semantic alignment (Li et al., 2024b; Xiong et al., 2025; Bachmann et al., 2025; Zheng et al., 2025b; Yao et al., 2025; Yu et al., 2025; Zheng et al., 2025a) , AR-aware encoding (Bachmann et al., 2025; Wang et al., 2025a; Li et al., 2024b; Miwa et al., 2025; Zhang et al., 2026; Wu et al., 2026; Zheng et al., 2025a), hierarchical/coarse-to-fine generation (Tian et al., 2024; Ren et al., 2024; Wang et al., 2025c; Tang et al., 2024; Huang et al., 2025; Liu et al., 2025; Zhang et al., 2026). But few of these works directly optimize the tokenizer to accommodate the AR model's actual distribution. The most direct way to achieve this is to jointly train an AR model and backpropagate its loss to the tokenizer. However, AR and reconstruction gradients share the same set of variables $\mathbf{z}_v$, placing the two objectives on a single-variable Pareto front: increasing λ degrades reconstruction, while decreasing λ weakens the AR training signal (Sec. 4.1). This limits how much generation can improve. We propose **Prologue**, which extends $\mathbf{z}_v$ to $[\mathbf{z}_p; \mathbf{z}_v]$, giving each objective its own variable: one group ($\mathbf{z}_p$) carries CE, while the other ($\mathbf{z}_v$) carries reconstruction. Prologue learns what information to encode in an image-dependent prefix through the AR objective, while keeping visual tokens dedicated to reconstruction.

Prologue is also grounded in a solid theoretical foundation. From the ELBO perspective, the prologue reduces prior matching from fitting the marginal $q(\mathbf{z}_v)$ to fitting the conditional $q(\mathbf{z}_v|\mathbf{z}_p)$, which is easier when $\mathbf{z}_p$ provides useful information. We further discuss how informative $\mathbf{z}_p$ can improve autoregressive modeling: since reconstruction objectives already preserve the information of visual tokens $\mathbf{z}_v$, collapsing $\mathbf{z}_p$ cannot improve reconstruction and removes image-dependent conditioning. Prologue tokens can instead capture information about $\mathbf{z}_v$ that lowers AR loss when the prediction benefit exceeds the additional prefix cost. Through AR training, prologue tokens can evolve into compact generation-oriented representations that improve autoregressive prediction while largely preserving reconstruction quality (Sec. 2.1).

Experimentally, Prologue significantly improves the quality of autoregressive visual generation, without relying on heuristic methods or auxiliary semantic regularization. On ImageNet 256×256, Prologue-Base reduces the gFID w/o CFG from 21.01 to 10.75 (↓48.8%), the gFID from 5.02 to 4.11 (↓18.1%), with reconstruction nearly unchanged (rFID 2.15 to 2.24) compared to the baseline. Scaling to Prologue-Large, it achieves competitive results with rFID of 0.99 and gFID of 1.46. Prologue tokens, driven purely by AR gradients, develop semantic structure (linear probe 35.88% vs. 23.71%); images resampled with fixed prologue tokens show similar overall structure with different visual details (Fig. 1). A natural variant, Prologue-Post, learns the prologue from a frozen visual tokenizer. It keeps rFID unchanged, and still reduces gFID vs. the 2D baseline by 22.7%: a prologue can be written together with the script, or composed separately after the script is complete.

2 PROLOGUE

As illustrated in Fig. 2, given an image $\mathbf{x}$, a shared Encoder E simultaneously receives learnable queries and patch embeddings:

$$[\mathbf{h}_p; \mathbf{h}_v] = E([\mathbf{q}_1, \ldots, \mathbf{q}_K; \text{Patchify}(\mathbf{x})]) \tag{1}$$

where $\mathbf{h}_p \in \mathbb{R}^{K\times d}$ corresponds to K prologue positions (default $K = 16$) and $\mathbf{h}_v \in \mathbb{R}^{N\times d}$ corresponds to $N = 256$ visual positions. The Encoder is a bidirectional Transformer with global attention shared across all positions.

The two groups are quantized through separate quantizers and codebooks: $\mathbf{h}_p$ is quantized by the Prologue Quantizer $\mathcal{Q}_p$ and Codebook $\mathcal{C}_p$ (size 1024) to yield $\mathbf{z}_p$, and $\mathbf{h}_v$ by the Visual Quantizer $\mathcal{Q}_v$ and Codebook $\mathcal{C}_v$ (size 16384 for Base and 4096 for Large) to yield $\mathbf{z}_v$. Here $\mathbf{z}_p$ and $\mathbf{z}_v$ are discrete ids produced by quantizers. Visual tokens are sent to the Decoder for image reconstruction: $\hat{\mathbf{x}} = D(\text{Lookup}(\mathcal{C}_v, \mathbf{z}_v))$, trained with ℓ_1, perceptual (Zhang et al., 2018), and adversarial losses (Isola et al., 2017). Prologue tokens do not participate in reconstruction.

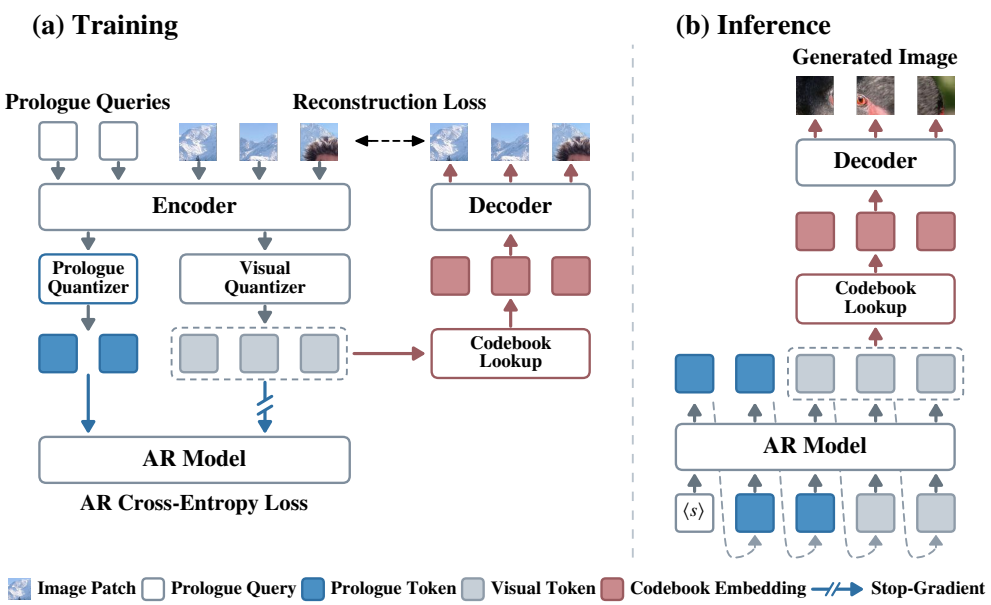


Figure 2: Prologue architecture. During training, a shared Encoder produces two groups of tokens: Visual Tokens (2D) handle reconstruction only, while Prologue Tokens (1D) receive AR gradients from the compact AR to reduce the data-model gap. The encoder is then frozen to train a full-size AR model (omitted for clarity). During inference, the full-size AR model performs standard autoregressive generation. The model will first generate prologue tokens, then generate visual tokens based on prologue tokens. Finally, the decoder maps visual tokens to the final image.

The two groups are concatenated as $[\mathbf{z}_p; \mathbf{z}_v]$ and fed into a compact AR Model. The total loss is:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{recon}}(\mathbf{x}, \hat{\mathbf{x}})}_{\text{acts on visual tokens only}} + \lambda \cdot \underbrace{\mathcal{L}_{\text{AR}}(\mathbf{z}_p, \mathbf{z}_v)}_{\text{acts on prologue tokens only}} \tag{2}$$

where $\mathcal{L}_{\text{AR}}$ is the autoregressive CE loss over the full sequence $[\mathbf{z}_p; \mathbf{z}_v]$. Prob-STE (Sec. 2.2) routes the gradient of $\mathcal{L}_{\text{AR}}$ to prologue tokens only; visual tokens do not directly receive AR gradients. AR gradients can indirectly influence visual representations through shared Encoder updates, but reconstruction changes little in our experiments (rFID rises by only 0.09 in Sec. 4.1). Prologue-Post (Sec. 2.3) freezes the entire visual pathway for complete isolation. From the perspective of visual tokens, the AR no longer needs to fit the difficult marginal distribution $q(\mathbf{z}_v)$ directly, but instead fits the simpler conditional distribution $q(\mathbf{z}_v|\mathbf{z}_p)$ under prologue conditioning.

2.1 Theoretical Analysis from the Evidence Lower Bound (ELBO) Perspective

As noted in the introduction, two-stage training is a special case of the ELBO under deterministic encoding (VQ). Let q_ϕ denote the encoder, p_ψ the decoder, and p_θ the AR prior. The VQ deterministic encoding $q_\phi(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{z} - \text{Enc}_\phi(\mathbf{x}))$ reduces the ELBO to:

$$\text{ELBO} = \underbrace{\log p_\psi(\mathbf{x}|\mathbf{z}^*)}_{\text{Stage 1: reconstruction}} + \underbrace{\log p_\theta(\mathbf{z}^*)}_{\text{Stage 2: AR}} \tag{3}$$

Two-stage training decouples these terms: q_ϕ is shaped for reconstruction alone and never considers the AR. Taking expectation over the dataset, the Stage 2 objective amounts to minimizing the cross-entropy $H(q_\phi, p_\theta) = H(q_\phi) + D_{\text{KL}}(q_\phi \| p_\theta)$. Optimizing θ can only shrink the D_{KL} term, while the structure of q_ϕ is determined entirely by Stage 1, which is the origin of the reconstruction-generation gap. We refer to $D_{\text{KL}}(q_\phi \| p_\theta)$ as the **data-model gap** hereafter.

ELBO of Prologue. Prologue augments the latent to $\mathbf{z} = [\mathbf{z}_p; \mathbf{z}_v]$, where $\mathbf{z}_p$ is an auxiliary latent that does not participate in reconstruction ($p_\psi(\mathbf{x}|\mathbf{z}_p, \mathbf{z}_v) = p_\psi(\mathbf{x}|\mathbf{z}_v)$) and only assists AR modeling



Figure 3: Samples from Prologue L–XL (semantic cfg=0.7, visual cosine cfg: scale=2.25, p=0.225).

through a structured prior. Under deterministic encoding:

$$\text{ELBO}_{\text{Prologue}} = \underbrace{\log p_\psi(\mathbf{x}|\mathbf{z}_v^*)}_{\text{reconstruction (same)}} + \underbrace{\log p_\theta(\mathbf{z}_p^*)}_{\text{prologue prior}} + \underbrace{\log p_\theta(\mathbf{z}_v^*|\mathbf{z}_p^*)}_{\text{conditional visual prior}} \tag{4}$$

Compared with Eq. 3, the reconstruction term is identical, while the prior term changes from $\log p_\theta(\mathbf{z}_v^*)$ to $\log p_\theta(\mathbf{z}_p^*) + \log p_\theta(\mathbf{z}_v^*|\mathbf{z}_p^*)$. The AR's prior matching shifts from fitting the marginal $q(\mathbf{z}_v)$ to fitting the conditional $q(\mathbf{z}_v|\mathbf{z}_p)$. For a fixed $q(\mathbf{z}_v)$, the inequality $H(\mathbf{z}_v|\mathbf{z}_p) \leq H(\mathbf{z}_v)$ implies a lower conditional entropy, provided that $\mathbf{z}_p$ carries non-trivial information about $\mathbf{z}_v$, i.e., $I(\mathbf{z}_p; \mathbf{z}_v) > 0$. We discuss below how informative prologue tokens can benefit AR prediction, while collapsed tokens provide no image-dependent information.

Learning an informative prologue. The loss $\mathcal{L}_{\text{recon}} + \lambda\mathcal{L}_{\text{AR}}$ Prologue optimizes is motivated by the ELBO in Eq. 4; gradient routing sends AR gradients only to $\mathbf{z}_p$, with reconstruction nearly unaffected (Sec. 4.1). With visual tokens trained for reconstruction only, $H(\mathbf{z}_v)$ can be treated as fixed in this analysis. For the unmasked AR loss, the remaining ϕ_p-dependent contributions decompose as $H(\mathbf{z}_p) - I(\mathbf{z}_p; \mathbf{z}_v)$ plus the remaining KL terms. Collapsing $\mathbf{z}_p$ to a constant zeros $H(\mathbf{z}_p)$ and $I(\mathbf{z}_p; \mathbf{z}_v)$ simultaneously, removing image-dependent conditioning. An informative prologue can achieve a lower AR loss than a collapsed one when its benefit to visual-token prediction exceeds its additional prediction cost. This motivates compressing information from $q(\mathbf{z}_v)$ into $q(\mathbf{z}_p)$ in a form the AR can predict autoregressively (full derivation in Append. C).

2.2 Implementation Details

Prob-STE. The AR CE loss is defined on discrete token indices. Standard STE performs straight-through estimation in the embedding space and is ineffective for discrete CE. Prob-STE first computes the soft assignment $p_i = \text{softmax}(\mathbf{h}_p \cdot \mathbf{c}_i/\tau)$; the forward pass uses $\arg\max$ to produce a one-hot vector, while the backward pass propagates gradients through $\tilde{p}_i = p_i + \text{sg}(\hat{p}_i - p_i)$. This is equivalent to the straight-through variant of Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) without Gumbel noise, an idea originating from an open-source implementation (Wang) and further scaled by IBQ (Shi et al., 2025). On the AR Model side, the input embedding for prologue tokens is computed as $\tilde{\mathbf{p}} \cdot \mathbf{W}_{\text{emb}}$ rather than a discrete lookup, so AR gradients can flow back to the Encoder along the soft probability (see more details in Append. E).

Visual token dropout. During Stage 1 training (Sec. 3.1), all visual tokens in the AR input are replaced with "[EOS]" with probability $p_{\text{drop}} = 0.5$. The visual part in Prologue contains 256 tokens, compared with only 16 prologue tokens, and the local dependencies among visual tokens are strong. The AR model may rely heavily on these visual tokens and ignore the prologue prefix, preventing prologue tokens from receiving useful gradients. Visual dropout mixes standard AR prediction with prediction from the prologue and masked visual inputs, while leaving the targets unchanged. $p_{\text{drop}} = 0.5$ strikes the best balance (Tab. A1).

CFG for Prologue. At inference time, different classifier-free guidance (Ho & Salimans, 2022) scales are applied to the two groups. Prologue tokens encode class-level information, so an excessively large scale reduces intra-class diversity; visual tokens encode spatial details and require stronger guidance. Model-specific guidance scales are listed in Tab. A6.

2.3 PROLOGUE-POST: A POST-HOC UPGRADE TO A FROZEN TOKENIZER.

As prologue tokens carry only AR gradients while visual tokens carry only reconstruction gradients, the two objectives are structurally decoupled by design, making the framework agnostic to *when* and *how* the prologue is introduced. At one end of this spectrum, we instantiate **Prologue-Post**, which attaches the prologue to an existing Tokenizer in a post-hoc manner: all parameters of the existing visual encoder, quantizer, and decoder are frozen, while only an additional prologue tokenizer $(\mathcal{E}_p, \mathcal{Q}_p)$ and a compact AR Model are newly introduced and trained with $\mathcal{L}_{\text{AR}}$. This allows us to learn a small set of prologue tokens for a frozen visual tokenizer and improve downstream AR generation (gFID ↓22.7% vs. the 2D baseline, Tab. A1), without retraining or modifying the original Tokenizer. The full-size AR model is still trained on the augmented token sequence; Prologue-Post reuses the tokenizer, not an already-trained AR model. We validate this variant on the Base configuration as a proof of concept (Sec. 4.3, Append. D), and leave large-scale exploration to future work.

3 EXPERIMENTS

3.1 SETUP

We conduct our main experiments on ImageNet (Deng et al., 2009) 256×256. We evaluate reconstruction results with rFID (Heusel et al., 2017) and PSNR, and class-conditional generation results with gFID, Inception Score (IS) (Salimans et al., 2016), Precision, and Recall. rFID is computed on the ImageNet validation set and gFID follows the ADM (Dhariwal & Nichol, 2021) protocol. The Tokenizer has B and L variants; the AR Model has B, L and X-Large variants. Prologue L–XL denotes a Large tokenizer paired with an XL AR model. We use a two-stage training pipeline. In Stage 1, the Tokenizer trains for 150/200 (B/L) epochs with a compact AR (40M) supplying AR gradients to prologue tokens (Sec. 2). In Stage 2, a full-size AR trains from scratch on the frozen tokenizer for 400/800 epochs when paired with the B/L tokenizer, with 150 additional epochs for L–XL. We also test a full one-stage training variant **Prologue-OneStage** but observe degraded results (Append. D). For more experimental details, see Append. E.

3.2 CONTROLLED COMPARISON

Tab. 1 compares the standard two-stage baseline with Prologue under the same experiment settings. Prologue significantly improves generation quality: gFID w/o CFG drops from 21.01 to 10.75 (↓48.8%) and gFID from 5.02 to 4.11 (↓18.1%), while reconstruction is almost the same (rFID 2.15/2.24, PSNR 20.71/20.64). Notably, Prologue-Post improves generation without affecting reconstruction at all. These results are consistent with our theoretical analysis (Sec. 2.1): prologue tokens can improve generation without compromising reconstruction. Prologue also outperforms static, class-conditioned, and global-pooled prefixes with and without CFG; static prefixes reach 15.59 gFID w/o CFG, compared with 10.75 for Prologue (Append. D.1). Under the matched L–B setting, Prologue improves gFID from 2.59 to 2.15 while keeping rFID at 0.99 (Tab. A3).

Table 1: Controlled comparison on ImageNet 256×256 (B–B; 86M Tokenizer, 86M Decoder, and 115M AR). Prologue uses $\lambda = 3$. We report the best gFID for each model.

Method	rFID ↓	PSNR ↑	gFID ↓	gFID w/o CFG ↓
1D Tokenizer	2.11	20.71	6.10	19.32
2D Tokenizer	2.15	20.71	5.02	21.01
Prologue	2.24	20.64	4.11	10.75
Prologue-Post	2.15	20.71	3.88	11.04

3.3 COMPREHENSIVE COMPARISON

Tab. 2 presents detailed reconstruction and generation comparisons. Prologue achieves competitive results against state-of-the-art methods. Specifically, Prologue-Base matches classical 2D-token AR baselines such as LlamaGen in compression rate and model size while improving generation quality (gFID 4.11 vs. 5.46 for LlamaGen-B). Prologue-Large adopts a compression rate and tokenizer size comparable to modern 1D-token AR works such as AliTok, and reaches comparable performance without their auxiliary training techniques (e.g., dual perceptual losses or periodic codebook clustering). Our best model, Prologue L–XL, reaches rFID 0.99 and gFID 1.46, on par with leading diffusion models, masked-prediction methods, and non-standard AR variants. We emphasize that Prologue is trained *without* the auxiliary techniques commonly used in recent

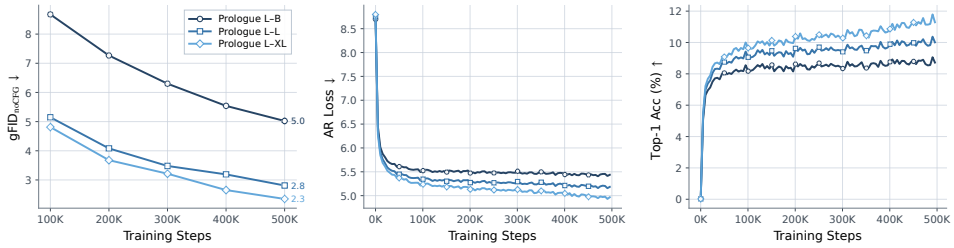


Figure 4: Training dynamics for Prologue L-series. **Left:** gFID w/o CFG. **Middle:** AR training loss. **Right:** Top-1 accuracy. Scaling the AR improves all three metrics throughout training.

Table 2: Comparison of reconstruction and class-conditional generation on ImageNet 256×256. †: Uses auxiliary semantic alignment to DINO/SigLIP features. ‡: Uses an additional periodic codebook/token clustering. §: Uses tail/nested/token dropout. ⋄: Uses a diffusion/flow-matching decoder on discrete tokens. ⋆: Distilled/Initialized from a pre-trained tokenizer. Prologue achieves competitive results without these techniques.

Method	Tokenizer				Generator					
	#Param(Enc+Dec).	Codebook	rFID↓	PSNR↑	#Param.	gFID↓	gFID$_{w/o\ CFG}$ ↓	IS↑	Pre.↑	Rec.↑
Continuous Tokens										
LDM-4-G (Rombach et al., 2022)	55M	-	0.27	27.53	400M	3.60	-	247.7	0.87	0.48
DiT-XL/2 (Peebles & Xie, 2023)	84M	-	0.62	-	675M	2.27	9.62	278.2	0.83	0.57
MAR-H (Li et al., 2024a)	70M	-	0.53	-	943M	1.55	2.35	303.7	0.81	0.62
REPA-SiT-XL† (Yu et al., 2025)	84M	-	0.62	-	675M	1.42	5.90	305.7	0.80	0.65
LightningDiT-XL† (Yao et al., 2025)	70M	-	0.28	-	675M	1.35	2.17	295.3	0.79	0.65
Discrete Tokens + Non-standard AR / Non-AR										
MAGVIT-v2 (Yu et al., 2024a)	-	262K	-	-	307M	1.78	3.65	319.4	-	-
VAR-*d*30 (Tian et al., 2024)	-	4K	-	-	2.0B	1.92	-	323.1	0.82	0.59
TiTok-B-64⋆ (Yu et al., 2024c)	86+86M	4K	1.70	-	177M	2.48	3.08	214.7	-	-
TiTok-S-128⋆ (Yu et al., 2024c)	22+22M	4K	1.71	-	287M	1.97	4.44	281.8	-	-
ImageFolder⋆†§ (Li et al., 2024b)	86+86M	4K×2	0.80	-	362M	2.60	-	295.0	0.75	0.63
RAR-B (Yu et al., 2024b)	36+36M	1K	2.28	-	261M	1.95	-	290.5	0.82	0.58
RAR-XXL (Yu et al., 2024b)	36+36M	1K	2.28	-	1.5B	1.48	-	326.0	0.80	0.63
ResTok†§ (Zhang et al., 2026)	662M	8K	1.28	-	326M	2.34	4.56	257.8	0.79	0.60
Discrete Tokens + Standard AR										
LlamaGen-B (Sun et al., 2024)	36+36M	16K	2.19	-	111M	5.46	26.26	193.6	0.83	0.45
LlamaGen-XXL (Sun et al., 2024)	36+36M	16K	2.19	-	1.4B	2.34	14.64	253.9	0.80	0.59
FlexTok d18-d28†§⋄ (Bachmann et al., 2025)	1.4B	64K	1.45	13.96	1.4B	1.71	-	284.9	0.82	0.61
GigaTok-B-L† (Xiong et al., 2025)	622M	16K	0.81	21.21	1.4B	2.03	-	238.5	0.80	0.63
VFMTok-SigLIP2-XXL† (Zheng et al., 2025b)	-	16K	0.94	-	1.4B	2.16	1.98	272.0	0.83	0.60
AliTok-B‡ (Wu et al., 2026)	390M	4K	0.86	20.97	177M	1.44	2.40	319.5	0.77	0.65
AliTok-XL‡ (Wu et al., 2026)	390M	4K	0.86	20.97	662M	1.28	1.88	306.3	0.79	0.65
Prologue (Ours, Discrete Tokens + Standard AR)										
Prologue B–B	86+86M	1K+16K	2.24	20.64	115M	4.11	10.75	210.3	0.83	0.48
Prologue B–L	86+86M	1K+16K	2.24	20.64	305M	2.67	6.56	251.2	0.82	0.56
Prologue B–XL	86+86M	1K+16K	2.24	20.64	685M	2.43	5.22	252.6	0.80	0.59
Prologue L–B	86+305M	1K+4K	0.99	20.01	115M	2.15	5.02	219.9	0.79	0.60
Prologue L–L	86+305M	1K+4K	0.99	20.01	305M	1.52	2.81	251.6	0.77	0.66
Prologue L–XL	86+305M	1K+4K	0.99	20.01	685M	1.46	2.26	257.7	0.78	0.66

tokenizers, including semantic alignment to DINO/SigLIP features, periodic codebook clustering, tail/nested token dropout, diffusion-based decoders, or distillation/initialization from pre-trained model. The tokenizer is trained from scratch with reconstruction, adversarial, and perceptual losses, and the generator is trained with a simple next-token prediction task.

3.4 Scalability

The bottom of Tab. 2 shows that Prologue scales well along both axes: the tokenizer and the AR model. Scaling the tokenizer from Base to Large reduces rFID from 2.24 to 0.99 and, under the same AR size, reduces gFID w/o CFG from 10.75 (B–B) to 5.02 (L–B). Scaling the AR model with the Base tokenizer, gFID w/o CFG drops from 10.75 (B–B) to 6.56 (B–L) and 5.22 (B–XL). With the Large tokenizer, gFID w/o CFG drops from 5.02 (L–B) to 2.81 (L–L) and 2.26 (L–XL). These results show consistent gains from scaling both the tokenizer and the AR model.

4 ANALYSIS

4.1 WHY PROLOGUE WORKS: λ SWEEP AND GRADIENT ANALYSIS

When AR gradients flow directly to visual tokens, $\nabla_\phi \mathcal{L}_{\text{AR}}$ includes $\nabla H(\mathbf{z}_v)$: the gradient can reduce CE by lowering the data entropy rather than improving AR fit. This phenomenon is observed across modalities (Wang et al., 2025b). Even with a very small λ, CE still alters the intrinsic structure of $\mathbf{z}_v$ (Sec. 4.2). Prologue structurally blocks this pathway: AR gradients cannot directly reach $H(\mathbf{z}_v)$. Collapsing $\mathbf{z}_p$ to a constant drives both $H(\mathbf{z}_p)$ and $I(\mathbf{z}_p; \mathbf{z}_v)$ to zero, thus providing no image-dependent conditioning. A useful optimization direction is for $\mathbf{z}_p$ to encode information that enables autoregressive prediction of $\mathbf{z}_v$ under the AR model, motivating an information bottleneck (full derivation in Append. C).

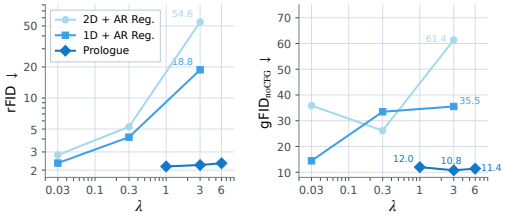


Figure 5: λ sweep comparison between 2D + AR Reg., 1D + AR Reg., and Prologue. **Left:** rFID vs. λ; **right:** gFID w/o CFG vs. λ. All models are trained with the B–B configuration with 86M Tokenizer, 86M Decoder, and 115M AR.

Fig. 5 illustrates this behavior by sweeping λ under three AR routing strategies. When reconstruction and AR gradients share $\mathbf{z}_v$, the two objectives trade off against each other and the trade-off worsens with λ. For reconstruction, rFID of both 1D and 2D tokenizers collapses at moderate λ (18.8 for 1D, 54.6 for 2D at $\lambda = 3$). Even $\lambda = 0.3$ already alters reconstruction, and only at $\lambda = 0.03$ does rFID stay close to the two-stage baseline and Prologue (but still worse). For generation, the 1D tokenizer benefits from joint AR training at small λ, but its best gFID w/o CFG (14.45) still lags Prologue (10.75). The 2D tokenizer behaves differently: joint AR training produces worse generation than the two-stage baseline. Our attention analysis (Sec. 4.2) shows that forcing 2D tokens into the 1D AR raster order destroys their original local attention pattern and makes AR fitting harder.

These results suggest that sharing a single latent for both reconstruction and generation couples the two objectives on a single-variable Pareto front, limiting how much generation can improve. After latent augmentation, the trade-off disappears: rFID fluctuates by only 0.16 as λ varies from 1 to 6; gFID w/o CFG reaches its minimum of 10.75 at $\lambda = 3$ (11.98 at $\lambda = 1$, 11.38 at $\lambda = 6$). Notably, Prologue operates at λ values (1–6) far above those feasible for direct AR (≤ 0.03), as the structural decoupling removes the reconstruction constraint on λ.

4.2 ATTENTION PATTERNS: THE STRUCTURAL ADVANTAGE OF 2D TOKENS

Tab. 3 and Fig. 6 compare 2D and 1D tokens. The 2D baseline's attention naturally exhibits a local pattern, with each token primarily attending to its left and top neighbors (Fig. 6a); this is a surprising intrinsic advantage of the 2D raster order. 1D tokens, produced by a bidirectional Encoder, lack this structure (Fig. 6b), making AR fitting harder (gFID of 1D: 6.10 vs. 2D: 5.02, Tab. 3).

Table 3: Comparison between 2D and 1D tokens. Under the same Encoder and AR sizes, 2D tokens show spatial locality, lower AR loss, and better generation with CFG. All models are trained under the B–B configuration with 86M Tokenizer, 86M Decoder, 115M AR, and without AR regularization.

Token structure	rFID ↓	PSNR ↑	AR Loss ↓	gFID ↓	gFID w/o CFG ↓
1D Tokenizer	2.11	20.71	7.89	6.10	19.32
2D Tokenizer	2.15	20.71	7.73	5.02	21.01

When AR gradients are applied to 2D visual tokens (Fig. 6c), the local structure is severely disrupted. The root cause is a *fundamental conflict* between the 2D spatial order encoded by the Tokenizer and the 1D raster order required by the AR model: AR gradients push visual tokens toward easier next-token prediction along the raster scan, destroying the spatial locality that made 2D tokens advantageous in the first place. Interestingly, this conflict is less severe for 1D tokens (Fig. 6d): Since 1D tokens carry no strong inherent ordering, AR gradients (or other supervision signals) can shape their order without overriding a pre-existing structure, an advantage of 1D tokenizers.

In Prologue, we use 1D tokens for the prologue tokens and 2D tokens for visual content to leverage the advantages of each in their corresponding role. The visual tokens in Prologue do not directly receive AR gradients, so the 2D local structure is fully preserved (Fig. 6). A new attention pattern

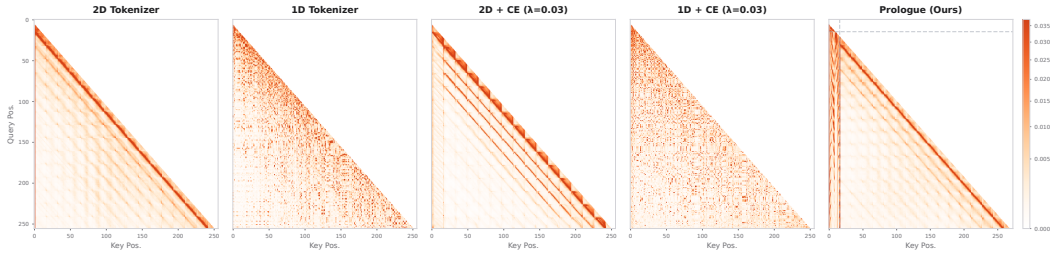


Figure 6: AR attention patterns (Layer 7; self-attention diagonal and first token masked). (a) 2D Tokenizer shows a clear local pattern (left neighbor + above). (b) 1D Tokenizer lacks spatial locality; attention is dispersed. (c) CE on 2D visual tokens ($\lambda = 0.03$) severely disrupts the local structure as the 2D spatial order inherently conflicts with the 1D AR raster order, and AR gradients destroy the spatial locality. (d) CE on 1D visual tokens ($\lambda = 0.03$) makes the attention pattern more uniform compared to the 1D baseline. (e) Prologue preserves the 2D local pattern while introducing prologue attention (vertical stripes; gray dashed line marks the prologue/visual boundary), showing that the AR model learns to use prologue tokens' global information to assist local visual token prediction. Full heatmaps across all layers are provided in Appendix Fig. A2.



Figure 7: Attention heatmaps showing visual-token attention to each prologue token (averaged from the last 8 layers and heads). Prologue tokens provide rich semantic information to visual tokens about objects and background in different regions, indicating a learned spatial division of labor.

also emerges: visual tokens attend heavily to the prologue prefix (vertical stripes on the left), showing that the AR Model indeed learns to use the global information from prologue tokens to assist local visual token prediction (Fig. 6e).

4.3 Ablation Study

We ablate Prologue design space under the B–B configuration: visual dropout p_{drop}, prologue token length K, prologue codebook size $|\mathcal{C}_p|$, Prob-STE temperature τ, AR Loss weight λ, Prologue-Post, and Prologue-OneStage. Visual dropout and prologue length affect generation quality, while codebook size and τ are relatively stable. Full results and discussion are in Append. D (Tab. A1).

4.4 Emergent Semantic Representations and Visualization

Tab. 4 uses linear probing on ImageNet classification (following MAE's protocol (He et al., 2022)) to investigate what information AR gradients drive prologue tokens to encode. We mean-pool the first 16 tokens produced by the Prologue tokenizer and the standard 2D tokenizer, and train a linear classifier on top. Prologue tokens reach 35.88% Top-1 accuracy, well above the 23.71% of the first 16 tokens from the 2D baseline (↑12.17 pp). These results show that AR-trained prologue tokens contain semantic information, consistent with the information bottleneck analysis (Sec. 4.1): prologue tokens naturally summarize image semantics under a pure generation objective.

Table 4: Linear probing results on ImageNet Classification task using only the *first 16 tokens* of each tokenizer. Driven solely by AR gradients, Prologue's 16 prologue tokens outperform the first 16 tokens of a standard 2D tokenizer.

Method	Top-1 ↑	Top-5 ↑
2D Tokenizer	23.71	44.17
Prologue	**35.88**	**60.28**

Fig. 1 illustrates the information division between the two groups through controlled variation. With prologue tokens held fixed and only visual tokens changed during generation, samples retain similar class identity and overall layout while texture and local details vary. Fig. 7 further visualizes the per-token attention: different prologue tokens specialize in different spatial regions (e.g., foreground

object vs. background), suggesting that different tokens capture different aspects of the global structure. These results suggest that prologue tokens provide global semantic conditioning under the AR objective.

5 Discussion

Prologue aims to supervise the tokenizer with the AR model's distribution. Existing approaches explore several directions to improve the AR-compatibility of tokens. One line aligns tokenizer features with pretrained vision models that carry dense semantic information, such as the DINO series (Oquab et al., 2024; Siméoni et al., 2025), to inject semantic structure into the tokens (Li et al., 2024b; Xiong et al., 2025; Bachmann et al., 2025; Zheng et al., 2025b; Yao et al., 2025; Yu et al., 2025; Zheng et al., 2025a). Another line introduces AR-aware regularization on the tokenizer, including causal decoders (Wu et al., 2026), tail-dropout objectives that implicitly induce sequential structure (Bachmann et al., 2025; Wang et al., 2025a; Li et al., 2024b; Miwa et al., 2025; Zhang et al., 2026), and flow-matching decoders (Wang et al., 2025a; Bachmann et al., 2025; Sargent et al., 2025). A third line designs hierarchical or coarse-to-fine tokenizers whose tokens align better with the AR prior (Tian et al., 2024; Ren et al., 2024; Wang et al., 2025c; Tang et al., 2024; Huang et al., 2025; Liu et al., 2025; Zhang et al., 2026). The most direct route is to backpropagate an AR loss to the tokenizer (Wang et al., 2025b; Ramanujan et al., 2025), but existing methods rely on very small AR weights or weaker surrogates (e.g., MSE) due to the entropy collapse problem we analyze in Sec. 4.1, which limits performance.

Prologue takes a different path: AR gradients are routed only to a dedicated set of prologue tokens, without directly supervising visual tokens. This avoids the collapse of visual tokens observed in prior direct-AR-loss approaches (Wang et al., 2025b; Ramanujan et al., 2025), and lets us close the data-model gap with the actual AR likelihood while largely preserving reconstruction quality. One could instead try to optimize $D_{\mathrm{KL}}(q_\phi \| p_\theta)$ directly. However, $H(q_\phi) = -\mathbb{E}_{q_\phi}[\log q_\phi]$ is a functional of the entire data distribution and cannot be computed per-sample for a deterministic encoder. Kernel discrepancy is feasible but depends on kernel choice and does not directly exploit the autoregressive structure of p_θ; score-based objectives require additional derivation in the discrete space. The cross-entropy naturally matches the per-token prediction structure of p_θ without additional design, making it the simplest viable path.

In the AR setting, Prologue realizes a form of learned conditioning through sequence prefixing: it keeps visual tokens dedicated to reconstruction while learning an additional prefix that reduces AR modeling difficulty. This idea can in principle extend to other generation paradigms (masked prediction, diffusion), though different conditioning mechanisms and gradient pathways would be needed; we leave this to future work.

We evaluate Prologue on ImageNet 256×256; larger-scale datasets and higher resolutions remain to be explored. The prologue token count $K = 16$ and codebook size 1024 work well in the current setting; the ablation (Sec. 4.3) shows that K has a large effect on gFID. The 16 prologue tokens add roughly 6% to the AR sequence length.

6 Conclusion

In this work, we propose **Prologue**, an approach to bridge the reconstruction-generation gap in autoregressive (AR) visual generation. During training, Prologue learns to generate a small set of prologue tokens prepended to the visual tokens; the prologue tokens are trained exclusively with the AR loss, while the visual tokens remain dedicated to the reconstruction objective. During generation, Prologue first generates the prologue tokens, then generates the visual tokens conditioned on them. Formally, this design conditions the AR prior as $p_\theta(\mathbf{z}_p)\, p_\theta(\mathbf{z}_v \mid \mathbf{z}_p)$, reducing modeling difficulty while protecting reconstruction quality. Prologue-Base reduces gFID w/o CFG from 21.01 to 10.75 (↓48.8%) and gFID from 5.02 to 4.11 (↓18.1%), with reconstruction nearly unchanged. Scaling to Prologue-Large, it achieves competitive results with rFID 0.99 and gFID 1.46 without additional tricks. Prologue tokens also naturally develop semantic structure under pure AR gradients: 16 tokens reach 35.88% Top-1 on linear probing, well above the 23.71% of a standard tokenizer. More broadly, this design principle of providing a dedicated latent variable for optimizing the data-model prior while keeping visual tokens dedicated to reconstruction can be applied to any system where a reconstruction-generation gap exists.

AI USE STATEMENT

LLMs were used for language polishing, coding assistance, reviewing mathematical formulas, and searching for related literature. The core method and research were developed without AI assistance. The authors take full responsibility for the final content of this paper.

REFERENCES

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2241–2292. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/bachmann25a.html.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20s.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. doi: 10.48550/ARXIV.2207.12598. URL https://doi.org/10.48550/arXiv.2207.12598.

Yuanhui Huang, Weiliang Chen, Wenzhao Zheng, Yueqi Duan, Jie Zhou, and Jiwen Lu. Spectralar: Spectral autoregressive visual generation, 2025. URL https://arxiv.org/abs/2506.10962.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

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, 2025. URL https://arxiv.org/abs/2501.07730.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11523–11532, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lee_Autoregressive_Image_Generation_Using_Residual_Quantization_CVPR_2022_paper.html.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, nov 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/66e226469f20625aaebddbe47f0ca997-Abstract-Conference.html.

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens, 2024b. URL https://arxiv.org/abs/2410.01756.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, aug 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Yiheng Liu, Liao Qu, Huichao Zhang, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Xian Li, Shuai Wang, Daniel K. Du, Fangmin Chen, Zehuan Yuan, and Xinglong Wu. Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction, 2025. URL https://arxiv.org/abs/2505.21473.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL http://arxiv.org/abs/1711.05101.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8ishA3LxN8.

Keita Miwa, Kento Sasaki, Hidehisa Arai, Tsubasa Takahashi, and Yu Yamaguchi. One-d-piece: Image tokenizer meets quality-controllable compression, 2025. URL https://arxiv.org/abs/2501.10064.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 45–55, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Pang_RandAR_Decoder-only_Autoregressive_Visual_Generation_in_Random_Orders_CVPR_2025_paper.html.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4055–4064. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/parmar18a.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Vivek Ramanujan, Kushal Tirumala, Armen Aghajanyan, Luke Zettlemoyer, and Ali Farhadi. When worse is better: Navigating the compression-generation tradeoff in visual tokenization, 2025. URL https://arxiv.org/abs/2412.16326.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Flowar: Scale-wise autoregressive image generation meets flow matching, 2024. URL https://arxiv.org/abs/2412.15205.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Kyle Sargent, Kyle Hsu, Justin Johnson, Li Fei-Fei, and Jiajun Wu. Flow to the mode: Mode-seeking diffusion autoencoders for state-of-the-art image tokenization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19471–19481, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Sargent_Flow_to_the_Mode_Mode-Seeking_Diffusion_Autoencoders_for_State-of-the-Art_Image_ICCV_2025_paper.html.

Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization, 2025. URL https://arxiv.org/abs/2412.02692.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024. URL https://arxiv.org/abs/2406.06525.

Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer, 2024. URL https://arxiv.org/abs/2410.10812.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9a24e284b187f662681440ba15c416fb-Abstract-Conference.html.

Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders, 2016. URL https://arxiv.org/abs/1606.05328.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6306–6315, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Bohan Wang, Zhongqi Yue, Fengda Zhang, Shuo Chen, Li'an Bi, Junzhe Zhang, Xue Song, Kennard Yanting Chan, Jiachun Pan, Weijia Wu, Mingze Zhou, Wang Lin, Kaihang Pan, Saining Zhang, Liyu Jia, Wentao Hu, Wei Zhao, and Hanwang Zhang. Selftok: Discrete visual tokens of autoregression, by diffusion, and for reasoning, 2025a. URL https://arxiv.org/abs/2505.07538.

Hanyu Wang, Saksham Suri, Yixuan Ren, Hao Chen, and Abhinav Shrivastava. Larp: Tokenizing videos with a learned autoregressive generative prior, 2025b. URL https://arxiv.org/abs/2410.21264.

Phil Wang. Vector quantize pytorch. https://github.com/lucidrains/vector-quantize-pytorch. URL https://github.com/lucidrains/vector-quantize-pytorch. GitHub repository.

Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, and Xihui Liu. Parallelized autoregressive visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12955–12965, June 2025c. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Parallelized_Autoregressive_Visual_Generation_CVPR_2025_paper.html.

Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. Maskbit: Embedding-free image generation via bit tokens. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=NYe2JuN3v3. Featured Certification, Reproducibility Certification.

Pingyu Wu, Kai Zhu, Yu Liu, Longxiang Tang, Jian Yang, Yansong Peng, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Towards sequence modeling alignment between tokenizer and autoregressive model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GT3obnZ5Fk.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation, 2025. URL https://arxiv.org/abs/2504.08736.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yao_Reconstruction_vs._Generation_Taming_Optimization_Dilemma_in_Latent_Diffusion_Models_CVPR_2025_paper.html.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=pfNyExj7z2.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022b. ISSN 2835-8856. URL https://openreview.net/forum?id=AFDcYJKhND. Featured Certification.

Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=gzqrANCF4g.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation, 2024b. URL https://arxiv.org/abs/2411.00776.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation, 2024c. URL https://arxiv.org/abs/2406.07550.

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, 2025. URL https://arxiv.org/abs/2410.06940.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 495–507, 2022. doi: 10.1109/TASLP.2021.3129994. URL https://doi.org/10.1109/TASLP.2021.3129994.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Xu Zhang, Cheng Da, Huan Yang, Kun Gai, Ming Lu, and Zhan Ma. Restok: Learning hierarchical residuals in 1d visual tokenizers for autoregressive image generation, 2026. URL https://arxiv.org/abs/2601.03955.

Anlin Zheng, Haochen Wang, Yucheng Zhao, Weipeng Deng, Tiancai Wang, Xiangyu Zhang, and Xiaojuan Qi. Hita: Holistic tokenizer for autoregressive image generation, 2025a. URL https://arxiv.org/abs/2507.02358.

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive image generation, 2025b. URL https://arxiv.org/abs/2507.08441.

Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric vqgan for stablediffusion, 2023. URL https://arxiv.org/abs/2306.04632.

A BROADER IMPACTS

Our work may positively impact research on efficient and controllable visual generation. The core idea of improving generative quality by decomposing prior matching into an easier conditional modeling problem can be applied to methods beyond discrete and autoregressive generation. This framework provides a way to improve generation quality without perturbing or altering the original data, thereby avoiding potential information loss and raising the upper bound of generative quality. Improvements in image generation quality may also introduce misuse risks, including misleading synthetic imagery, impersonation, and disinformation. We therefore recommend that any release of checkpoints or generated samples be accompanied by clear usage guidelines, model cards, provenance documentation, and, where appropriate, watermarking or detection tools.

B RELATED WORK

Discrete image tokenization. The paradigm of discrete image tokens was established by VQ-VAE (van den Oord et al., 2017); VQGAN (Esser et al., 2021) introduced adversarial training and markedly improved reconstruction quality. Subsequent work has improved codebook utilization and efficiency through lookup-free quantization (LFQ) (Yu et al., 2024a), finite scalar quantization (FSQ) (Mentzer et al., 2024), residual quantization (RQ) (Lee et al., 2022). On the reconstruction frontier, improved VQGAN variants (Yu et al., 2022a; Zhu et al., 2023) and mode-seeking diffusion autoencoders (FlowMo) (Sargent et al., 2025) have pushed rFID substantially lower. On the token structure side, TiTok (Yu et al., 2024c) introduced a compact 1D tokenizer using learnable queries, FlexTok (Bachmann et al., 2025) supports variable-length encoding, and One-D-Piece (Miwa et al., 2025) explores quality-controllable compression. ImageFolder (Li et al., 2024b) uses semantic-detail dual-branch product quantization, and HART (Tang et al., 2024) decomposes a continuous latent into discrete and continuous parts.

Bridging the tokenizer–AR gap. A central challenge in two-stage pipelines is that tokens optimized for reconstruction may be unnecessarily difficult for the AR prior to model. Existing work explores several orthogonal directions to improve AR-compatibility. *Semantic alignment:* GigaTok (Xiong et al., 2025) aligns decoder features with frozen DINOv2 (Oquab et al., 2024); ResTok (Zhang et al., 2026) matches DINOv3 (Siméoni et al., 2025) features at both encoder and decoder; VFMTok (Zheng et al., 2025b) builds the entire encoder on top of frozen vision foundation models; ImageFolder (Li et al., 2024b) uses DINOv2 contrastive losses in a dual-branch tokenizer. *AR-aware encoding:* AliTok (Wu et al., 2026) trains a causal decoder to align with the AR generation order; tail-dropout objectives, first explored in neural audio codecs (Zeghidour et al., 2022) and later adapted to visual tokenizers (Miwa et al., 2025; Bachmann et al., 2025; Li et al., 2024b; Wang et al., 2025a), implicitly induce sequential structure by requiring reconstruction from token prefixes. *Hierarchical and coarse-to-fine generation:* VAR (Tian et al., 2024) predicts at progressively finer scales, FlowAR (Ren et al., 2024) integrates flow matching across scales, ResTok (Zhang et al., 2026) uses hierarchical residuals so that lower-level tokens carry global information, and PAR (Wang et al., 2025c) parallelizes generation across spatially distant tokens. *Direct generative-loss backpropagation:* The most natural approach is to backpropagate the AR loss directly to the tokenizer (Wang et al., 2025b; Ramanujan et al., 2025). For deterministic encoding, $H(q_\phi(\mathbf{z}|\mathbf{x})) = 0$, while the dataset-average AR cross-entropy contains the aggregate token entropy $H(q_\phi(\mathbf{z}))$. Reducing this entropy provides a shortcut to lower AR loss and can drive the tokenizer toward degenerate solutions. Existing mitigations rely on reduced loss weights, surrogate MSE objectives, limiting how much generation can improve (Sec. 4.1). Prologue retains CE as the optimization objective but routes the gradients to dedicated prologue tokens, structurally decoupling the two objectives. It is orthogonal to the semantic alignment and order-aware directions above and can in principle be combined with them.

Autoregressive visual generation. Autoregressive image generation dates back to PixelCNN-style decoders (van den Oord et al., 2016) and has progressed to Transformer-based systems (Parmar et al., 2018; Chen et al., 2020; Ramesh et al., 2021; Yu et al., 2022b). Building on VQ-based AR pipelines (van den Oord et al., 2017; Esser et al., 2021), LlamaGen (Sun et al., 2024) demonstrated that standard LLM-style backbones can scale autoregressive image generation to competitive quality. Beyond raster-order next-token prediction, recent work explores diverse generation paradigms: VAR (Tian et al., 2024) uses next-scale prediction, FlowAR (Ren et al., 2024) integrates flow matching,

HART (Tang et al., 2024) couples discrete AR with a residual diffusion module, DetailFlow (Liu et al., 2025) and SpectralAR (Huang et al., 2025) re-order tokens along resolution or spectral axes, RandAR (Pang et al., 2025) and RAR (Yu et al., 2024b) randomize the generation order. On the continuous side, MAR (Li et al., 2024a) removes vector quantization entirely and uses a diffusion head for per-token generation.

Relation to prefix tuning and prompt tuning. Prologue prepends a learned token prefix to the visual sequence, which bears a superficial resemblance to prefix tuning (Li & Liang, 2021) and prompt tuning (Lester et al., 2021) in NLP. However, the two ideas differ fundamentally: prefix/prompt tuning attaches *static* learned vectors that are shared across all inputs and independent of any specific instance, serving as a task-level conditioning mechanism. In contrast, Prologue learns a *tokenizer* that extracts instance-specific information from each image and compresses it into a compact prefix optimized for autoregressive prediction. In other words, prefix tuning conditions on a task; Prologue conditions on the data.

C JOINT CROSS-ENTROPY DECOMPOSITION

Prologue transforms the AR modeling objective from marginal matching $p_\theta(\mathbf{z}_v) \approx q(\mathbf{z}_v)$ to conditional matching $p_\theta(\mathbf{z}_v|\mathbf{z}_p) \approx q(\mathbf{z}_v|\mathbf{z}_p)$. Below we derive the mathematical properties of this transformation starting from the CE of the joint distribution. Let $q_\phi(\mathbf{z}_p, \mathbf{z}_v)$ denote the joint data prior induced by the Encoder, and $p_\theta(\mathbf{z}_p, \mathbf{z}_v)$ the joint model prior from the AR. The joint CE decomposes as:

$$\mathcal{L}_{\text{CE}} = H(q_\phi(\mathbf{z}_p, \mathbf{z}_v)) + D_{\text{KL}}(q_\phi(\mathbf{z}_p, \mathbf{z}_v) \| p_\theta(\mathbf{z}_p, \mathbf{z}_v)) \tag{5}$$

The data prior entropy decomposes as:

$$H(\mathbf{z}_p, \mathbf{z}_v) = H(\mathbf{z}_v) + H(\mathbf{z}_p|\mathbf{z}_v) = H(\mathbf{z}_v) + H(\mathbf{z}_p) - I(\mathbf{z}_p; \mathbf{z}_v) \tag{6}$$

With visual tokens trained for reconstruction only, $H(\mathbf{z}_v)$ can be treated as fixed in this analysis. Under this approximation, the unmasked CE objective for ϕ_p (prologue encoder) is:

$$\min_{\phi_p} \left[H(\mathbf{z}_p) - I(\mathbf{z}_p; \mathbf{z}_v) + D_{\text{KL}}(q_\phi(\mathbf{z}_p) \| p_\theta(\mathbf{z}_p)) + \mathbb{E}_{q(\mathbf{z}_p)}[D_{\text{KL}}(q_\phi(\mathbf{z}_v|\mathbf{z}_p) \| p_\theta(\mathbf{z}_v|\mathbf{z}_p))] \right] \tag{7}$$

Collapse pathway analysis. If $\mathbf{z}_p$ degenerates to a constant c ($H(\mathbf{z}_p) \to 0$), then $I(\mathbf{z}_p; \mathbf{z}_v) \to 0$. As the prior learns the constant prefix, $D_{\text{KL}}(q(\mathbf{z}_p) \| p_\theta(\mathbf{z}_p)) \to 0$, and the remaining visual KL is $D_{\text{KL}}(q(\mathbf{z}_v) \| p_\theta(\mathbf{z}_v|c))$. The total CE then approaches $H(\mathbf{z}_v) + D_{\text{KL}}(q(\mathbf{z}_v) \| p_\theta(\mathbf{z}_v|c))$, recovering the baseline form of visual-token modeling without image-dependent prefix information.

Effective pathway analysis. Informative $\mathbf{z}_p$ reduces $H(\mathbf{z}_v|\mathbf{z}_p)$ by increasing $I(\mathbf{z}_p; \mathbf{z}_v)$ and can help the AR fit the conditional distribution. The increase in $H(\mathbf{z}_p)$ incurs a prediction cost. With a compact prefix (16 tokens, 1K codebook), the aim is for the benefit from conditioning to exceed this additional prediction cost. AR training can thus encourage prologue tokens to encode global information that helps predict visual tokens, forming an information bottleneck.

Contrast: why direct optimization of visual tokens collapses. If AR gradients flow directly to ϕ_v (visual encoder), $\nabla_{\phi_v} \mathcal{L}_{\text{CE}}$ contains $\nabla_{\phi_v} H(\mathbf{z}_v)$, and the gradient can push $H(\mathbf{z}_v)$ to 0 (codebook degenerates to a constant). This shortcut substantially reduces CE but collapses reconstruction. Prologue structurally blocks this shortcut: AR gradients cannot directly reach $H(\mathbf{z}_v)$.

D ABLATION STUDY

Tab. A1 varies one default setting at a time ($p_{\text{drop}} = 0.5$, $K = 16$, $|\mathcal{C}_p| = 1024$, $\tau = 0.1$, $\lambda = 3.0$).

Visual dropout p. At $p = 0.0$, the AR Model can drive CE very low using only the local patterns of visual tokens, drowning out the prologue gradient signal; gFID reaches only 4.28. At $p = 1.0$, prologue tokens receive the strongest gradient, but the AR Model never sees unmasked visual inputs during Stage 1. $p = 0.5$ is the best balance.

Table A1: Ablation study (B–B). Default configuration highlighted.

Factor	Value	rFID ↓	gFID w/o CFG ↓	gFID ↓
Default	–	2.24	10.75	4.11
Prologue codebook	256	2.22	12.10	4.36
	4096	2.25	12.68	4.36
Prologue token count	4	2.24	13.49	4.20
	64	2.25	12.31	6.81
Prob-STE τ	0.01	2.18	11.66	4.02
	1.0	2.14	11.49	4.08
AR weight λ	1.0	2.17	11.98	4.21
	6.0	2.33	11.38	4.28
Visual dropout p	0.0	2.16	12.32	4.28
	1.0	2.25	12.67	4.53
Training Prologue with Pre-trained Visual Tokenizer				
Prologue-Post	–	2.15	11.04	3.88
Training Prologue with OneStage				
Prologue-OneStage	–	2.09	21.00	5.41

Prologue token count K and codebook size. At $K = 4$, the bottleneck is too tight and gFID reaches 4.20; at $K = 64$, the sequence is too long and gFID degrades to 6.81. Codebook size is relatively insensitive: from 256 to 4096, gFID is 4.36 in both cases.

Prob-STE τ and λ. τ is insensitive within a reasonable range (0.01 to 1.0, gFID 4.02 to 4.11). λ from 1.0 to 6.0 causes rFID to rise by only 0.16, whereas CE directly on visual tokens at comparable λ severely degrades reconstruction (Fig. 5).

Prologue-Post. As described in Sec. 2.3, Prologue-Post freezes the entire pre-trained visual pathway and introduces only the prologue components. rFID stays exactly at 2.15, with gFID 3.88 and gFID w/o CFG 11.04 under the same search protocols as Tab. 1; the default jointly trained model obtains gFID 4.11 and rFID 2.24 with shared Encoder updates.

Prologue-OneStage. We also evaluate a simplified one-stage training pipeline in which the full-size AR model is trained jointly with the tokenizer, eliminating the separate Stage 2. We observe that this leads to degraded generation quality compared to two-stage training, which is expected since the tokenizer's latent distribution shifts continuously during joint optimization and is not determined until the tokenizer is frozen.

Large Tokenizer variants and end-to-end generation quality. Recent tokenizers have explored stronger LPIPS backbones: MaskBit (Weber et al., 2024) uses ResNet-50, while TA-TiTok (Kim et al., 2025) and AliTok (Wu et al., 2026) use ConvNeXt-S (alongside VGG in AliTok). Following this trend, we replace VGG16 (Simonyan & Zisserman, 2015) with a standalone ConvNeXt-S (Liu et al., 2022) in our larger tokenizer. As shown in Table A2, this reduces rFID from 1.88 to 0.99 (↓47.3%) at 4K codebook size, while VGG16 retains an edge on PSNR, indicating that the two losses optimize along different axes. Expanding the codebook from 4K to 16K further lowers rFID to 0.92, but gFID slightly worsens (2.15 to 2.30), suggesting larger codebooks improve reconstruction precision at the cost of AR modeling difficulty. Prologue-Base retains VGG to match classical baselines such as LlamaGen, while Prologue-Large adopts ConvNeXt-S to match modern tokenizers such as AliTok. The improvement from ConvNeXt-S is orthogonal to Prologue itself: Table A3 shows Prologue provides consistent gains at both backbones.

Prologue improvement over the 2D baseline. Tab. A3 compares the standard 2D Tokenizer and Prologue at two scales, varying only whether prologue tokens are introduced. All other factors (Tokenizer size, perceptual loss, codebook, AR Model, training steps) are identical within each group.

Table A2: Effect of perceptual loss backbone and codebook size (L–B). All configurations use the same architecture and training hyperparameters ($\lambda = 3$, $\tau = 0.1$, $p_{\text{drop}} = 0.5$). ConvNeXt-S improves rFID and gFID over VGG16 at 4K. Reducing the codebook from 16K to 4K slightly increases rFID but improves gFID.

Perceptual loss	CB	BPP	rFID ↓	PSNR ↑	gFID w/o CFG ↓	gFID ↓
VGG16	4K	0.049	1.88	**20.45**	10.87	4.56
ConvNeXt-S	4K	0.049	0.99	20.01	5.02	**2.15**
ConvNeXt-S	16K	0.057	**0.92**	20.36	6.73	2.30

Table A3: Prologue improvement over the 2D baseline. The only variable within each group is whether prologue tokens are introduced; all other factors are identical.

Scale	Method	rFID ↓	gFID w/o CFG ↓	Δ	gFID ↓	Δ
B–B	2D Baseline	2.15	21.01	–	5.02	–
	Prologue	2.24	10.75	↓48.8%	4.11	↓18.1%
L–B	2D Baseline	0.99	7.54	–	2.59	–
	Prologue	0.99	5.02	↓33.4%	2.15	↓17.0%

D.1 Comparison with simpler prefixes

We compare three simpler 16-token prefixes with Prologue under B–B. Static prefixes use 16 learnable embeddings shared across images. Class-conditioned prefixes repeat the corresponding embedding from a table of 1,000 class embeddings 16 times. Global-pooled prefixes average-pool the original visual-token embeddings using 4×4 spatial pooling, map each pooled vector to its nearest codebook entry, and follow raster order. The AR model generates these 16 tokens before the visual tokens. Static and class-conditioned prefixes are trained with the AR loss; global-pooled tokens are not. Prologue tokens are image-dependent and trained with the AR loss. We search sampling parameters separately for each variant and report its best gFID.

Table A4: Comparison with simpler prefixes on ImageNet 256×256 (B–B, $K = 16$).

Prefix	Depends on	AR-shaped	rFID ↓	gFID w/o CFG ↓	gFID ↓
None (2D baseline)	–	–	2.15	21.01	5.02
Static learnable	–	Yes	2.15	15.59	4.86
Class-conditioned	Class label	Yes	2.15	14.65	17.46
Global-pooled image tokens	Image	No	2.15	17.55	5.02
Prologue	Image	Yes	2.24	10.75	4.11

Tab. A4 shows that all three simpler prefixes improve generation without CFG. Static prefixes also slightly improve gFID with CFG, while global pooling leaves it unchanged and class-conditioned prefixes perform worse in this setting. Prologue achieves the best gFID both with and without CFG. These results show that simpler prefixes recover part of the gain, but none matches the complete image-dependent and AR-shaped design across both settings.

E Implementation details

Computational resources. All experiments are conducted on a cluster of NVIDIA GPUs. We report the estimated training cost converted to an A100 setting. For Stage 1, the Base and Large tokenizers require approximately 30 and 44 hours on 8 A100 GPUs, respectively. For Stage 2, the Base and Large AR models require approximately 13 and 25 hours on 8 A100 GPUs, respectively, while the XL AR model requires approximately 25 hours on 32 A100 GPUs.

Training and inference overhead. Relative to the 2D baseline, Prologue adds 16 prologue tokens and one EOS token to 256 visual tokens, an increase of approximately 6.6%. The compact AR used in Stage 1 has 40M parameters. Stage 1 and Stage 2 training are 23.8% and 16.0% slower per iteration,

respectively. Inference throughput decreases from 153.8 to 147.1 images per second, a reduction of 4.4%.

Codebook utilization. The visual codebook of Prologue-Base reaches 100% utilization, with approximately 13.9 bits of aggregated posterior entropy out of a 14-bit maximum. For the Large configuration, the visual-codebook entropy is 11.53 bits out of a 12-bit maximum. The prologue codebook has lower per-position utilization, at 38.6%, with approximately 8.7 bits of entropy out of a 10-bit maximum. We do not use periodic codebook clustering.

Model configurations. Tab. A5 summarizes the architecture details of all model variants used in our experiments. The tokenizer consists of a ViT-based encoder and decoder, offered in Base (B) and Large (L) configurations. The standalone AR model is a decoder-only Transformer available in three sizes: Base (115M), Large (305M), and XL (685M). We denote each full system by its tokenizer–AR pairing, e.g., Prologue B–L refers to a Base tokenizer paired with a Large AR model.

Table A5: Model configurations. Tokenizer: Encoder+Decoder; AR: decoder-only Transformer.

	Params	Layers	Dim	Heads
Tokenizer (Encoder/Decoder)				
B	86M/86M	12/12	768/768	12
L	86M/305M	12/24	768/1024	12/16
AR Model				
B	115M	12	768	12
L	305M	24	1024	16
XL	685M	32	1280	20

Training details **Stage 1 (Tokenizer + joint AR)**. The encoder, decoder, and GAN discriminator are optimized with AdamW (Loshchilov & Hutter, 2017) ($\beta_1 = 0$, $\beta_2 = 0.99$), while the compact AR head uses AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is 1×10^{-4}, with linear warmup followed by linear decay to 3×10^{-5} for the Base model and 1×10^{-5} for the Large model. We use a batch size of 256 and train for 150 epochs (Base) or 200 epochs (Large). We use vanilla L2-Normed VQ (Yu et al., 2022a) for visual quantizer and Prob-STE for prologue quantizer. The loss weights are set to $\ell_1 = 1$, LPIPS $= 1$, GAN$_G = 0.1$, $\lambda_{\text{CE}} = 3$, and commitment $= 1$. No weight decay is applied to the encoder and decoder, while the compact AR uses a weight decay of 3×10^{-2}. Data augmentation consists of random cropping and horizontal flipping. **Stage 2 (Standalone AR training)**. The AR model paired with the Base tokenizer is trained with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.96$) at a learning rate of 2×10^{-4} with linear decay to 5×10^{-5}, a batch size of 512, for 400 epochs. When paired with the Large tokenizer, we increase the learning rate to 4×10^{-4} with linear decay to 1×10^{-5}, the batch size to 2048, and train for 800 epochs. Both configurations use a dropout rate of 0.1, an embedding dropout rate of 0.1, gradient clipping at 1.0, and a weight decay of 3×10^{-2}. For Prologue L–XL, we apply a weight decay of 5×10^{-2} and extend training by an additional 150 epochs, as the AR model has not fully converged after 800 epochs. The class label is dropped with 10% probability for classifier-free guidance. Data augmentation uses ten-crop.

Prob-STE details. Given the Encoder output $\mathbf{h}_p$ and codebook vectors $\{\mathbf{c}_i\}$, Prob-STE computes the soft assignment $p_i = \exp(\mathbf{h}_p \cdot \mathbf{c}_i/\tau)/\sum_j \exp(\mathbf{h}_p \cdot \mathbf{c}_j/\tau)$, where τ is the temperature. The forward pass uses hard assignment $\hat{p}_i = \mathbf{1}[i = \arg\max_j p_j]$; the backward pass propagates gradients through $\tilde{p}_i = p_i + \text{sg}(\hat{p}_i - p_i)$. The quantized representation is $\tilde{\mathbf{q}}_p = \sum_i \tilde{p}_i \cdot \mathbf{c}_i$.

Prologue and visual tokens use standard CE with the same treatment of targets. Prob-STE applies only to the prologue input embeddings. Visual dropout changes only the inputs; the targets remain unchanged. The CE loss uses mean reduction over all predicted tokens.

gFID / IS trend across the CFG scale. Fig. A1 shows how gFID and IS vary with the visual CFG scale s_{vis} for all six Prologue variants. For each variant we fix the semantic scale s_{pro} at its per-variant optimum, set the cosine shape parameter to $p = 0.25$, and sweep s_{vis} over the CFG evaluation set. Every variant exhibits the same trade-off: gFID traces a clean U-shape with an interior minimum,

while IS increases monotonically with s_{vis}—stronger visual guidance yields more class-confident samples at the cost of reduced diversity, and the best gFID is reached before IS saturates. The CFG configurations used for the reported models are listed in Tab. A6.

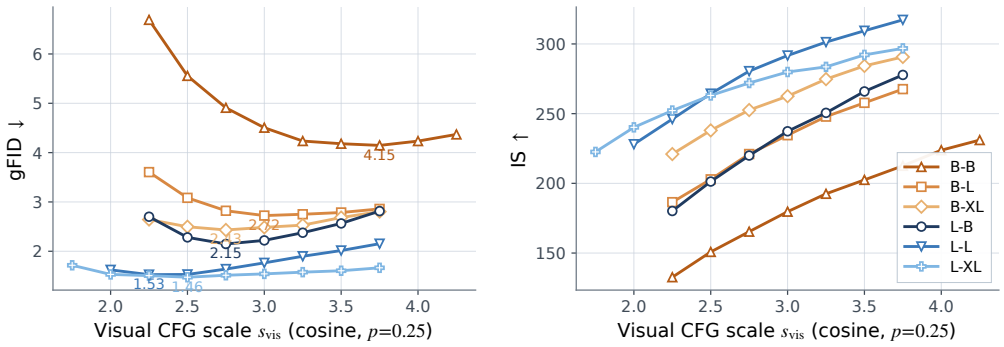


Figure A1: Trend of gFID (left) and IS (right) as the visual CFG scale s_{vis} varies (cosine schedule, $p = 0.25$). Each curve fixes s_{pro} at the per-variant optimum. gFID forms an interior-minimum U-shape for all six variants while IS increases monotonically, reflecting the fidelity–diversity trade-off.

Detailed CFG configuration. We provide the detailed sampling configurations—including classifier-free guidance scales and temperature settings—for the main models and selected ablations in Tab. A6.

F Additional Visualization and Results.

This appendix provides additional qualitative results. Fig. A2 visualizes AR attention patterns across layers. Fig. A3 shows semantic-fix generations, where prologue tokens are held fixed while visual tokens are resampled. Fig. A4 presents per-prologue-token attention heatmaps averaged over the last layers. Finally, Figs. A5–A9 and Figs. A10–A14 show uncurated random samples from Prologue L–XL and the 2D Tokenizer B–B baseline, respectively, generated without manual selection using non-overlapping random ImageNet classes.

Table A6: Sampling configurations used for the reported gFID.

Model	Mode	Prologue CFG	Visual CFG	Temperature
1D Tokenizer	CFG	–	cosine, $p = 1.75$, scale $= 19$	$T = 1.0$
1D Tokenizer	w/o CFG	–	–	$T = 0.85$
2D Tokenizer	CFG	–	cosine, $p = 1.5$, scale $= 10$	$T = 1.0$
2D Tokenizer	w/o CFG	–	–	$T = 0.85$
2D Tokenizer L–B	CFG	–	cosine, $p = 1.75$, scale $= 5$	$T = 1.0$
2D Tokenizer L–B	w/o CFG	–	–	$T = 0.9$
Prologue B–B	CFG	constant scale $= 0.7$	cosine, $p = 0.2$, scale $= 3.75$	$T = 1.0$
Prologue B–B	w/o CFG	–	–	$T_{\text{prologue}} = 0.7$, $T_{\text{visual}} = 0.9$
Prologue-Post	CFG	constant scale $= 0.6$	cosine, $p = 0.25$, scale $= 3.75$	$T = 1.0$
Prologue-Post	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.8$
Prologue-OneStage	CFG	constant scale $= 0.8$	cosine, $p = 0.25$, scale $= 3.75$	$T = 1.0$
Prologue-OneStage	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.8$
Prologue B–L	CFG	constant scale $= 0.8$	cosine, $p = 0.225$, scale $= 3.0$	$T = 1.0$
Prologue B–L	w/o CFG	–	–	$T_{\text{prologue}} = 0.7$, $T_{\text{visual}} = 0.9$
Prologue B–XL	CFG	constant scale $= 0.8$	cosine, $p = 0.25$, scale $= 2.75$	$T = 1.0$
Prologue B–XL	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.9$
Prologue L–B / ConvNeXt-S 4K	CFG	constant scale $= 0.65$	cosine, $p = 0.25$, scale $= 2.75$	$T = 1.0$
Prologue L–B / ConvNeXt-S 4K	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.95$
Prologue L–L	CFG	constant scale $= 0.7$	cosine, $p = 0.25$, scale $= 2.5$	$T = 1.0$
Prologue L–L	w/o CFG	–	–	$T_{\text{prologue}} = 0.95$, $T_{\text{visual}} = 0.9$
Prologue L–XL	CFG	constant scale $= 0.7$	cosine, $p = 0.225$, scale $= 2.25$	$T = 1.0$
Prologue L–XL	w/o CFG	–	–	$T_{\text{prologue}} = 1.0$, $T_{\text{visual}} = 0.9$
Prologue L–B / VGG16 4K	CFG	constant scale $= 0.7$	cosine, $p = 0.25$, scale $= 3.75$	$T = 1.0$
Prologue L–B / VGG16 4K	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.85$
Prologue L–B / ConvNeXt-S 16K	CFG	constant scale $= 0.65$	cosine, $p = 0.25$, scale $= 3.25$	$T = 1.0$
Prologue L–B / ConvNeXt-S 16K	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.95$
Prologue codebook 256	CFG	constant scale $= 0.7$	cosine, $p = 0.25$, scale $= 3.25$	$T = 1.0$
Prologue codebook 256	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.9$
Prologue codebook 4096	CFG	constant scale $= 0.8$	cosine, $p = 0.25$, scale $= 3.5$	$T = 1.0$
Prologue codebook 4096	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.9$
Prologue token count $K = 4$	CFG	constant scale $= 0.6$	cosine, $p = 0.25$, scale $= 2.5$	$T = 1.0$
Prologue token count $K = 4$	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.9$
Prologue token count $K = 64$	CFG	constant scale $= 0.8$	cosine, $p = 0.25$, scale $= 3.75$	$T = 1.0$
Prologue token count $K = 64$	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 1.0$
Prologue Prob-STE $\tau = 0.01$	CFG	constant scale $= 0.65$	cosine, $p = 0.25$, scale $= 2.5$	$T = 1.0$
Prologue Prob-STE $\tau = 0.01$	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.9$
Prologue Prob-STE $\tau = 1.0$	CFG	constant scale $= 0.8$	cosine, $p = 0.25$, scale $= 3.75$	$T = 1.0$
Prologue Prob-STE $\tau = 1.0$	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.85$
Prologue Visual dropout $p = 0.0$	CFG	constant scale $= 0.65$	cosine, $p = 0.25$, scale $= 3.5$	$T = 1.0$
Prologue Visual dropout $p = 0.0$	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.85$
Prologue Visual dropout $p = 1.0$	CFG	constant scale $= 0.75$	cosine, $p = 0.25$, scale $= 3.25$	$T = 1.0$
Prologue Visual dropout $p = 1.0$	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.9$
Prologue $\lambda = 1.0$	CFG	constant scale $= 0.65$	cosine, $p = 0.25$, scale $= 3.75$	$T = 1.0$
Prologue $\lambda = 1.0$	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.85$
Prologue $\lambda = 3.0$	CFG	constant scale $= 0.7$	cosine, $p = 0.2$, scale $= 3.75$	$T = 1.0$
Prologue $\lambda = 3.0$	w/o CFG	–	–	$T_{\text{prologue}} = 0.7$, $T_{\text{visual}} = 0.9$
Prologue $\lambda = 6.0$	CFG	constant scale $= 0.75$	cosine, $p = 0.25$, scale $= 3.25$	$T = 1.0$
Prologue $\lambda = 6.0$	w/o CFG	–	–	$T_{\text{prologue}} = 0.8$, $T_{\text{visual}} = 0.9$

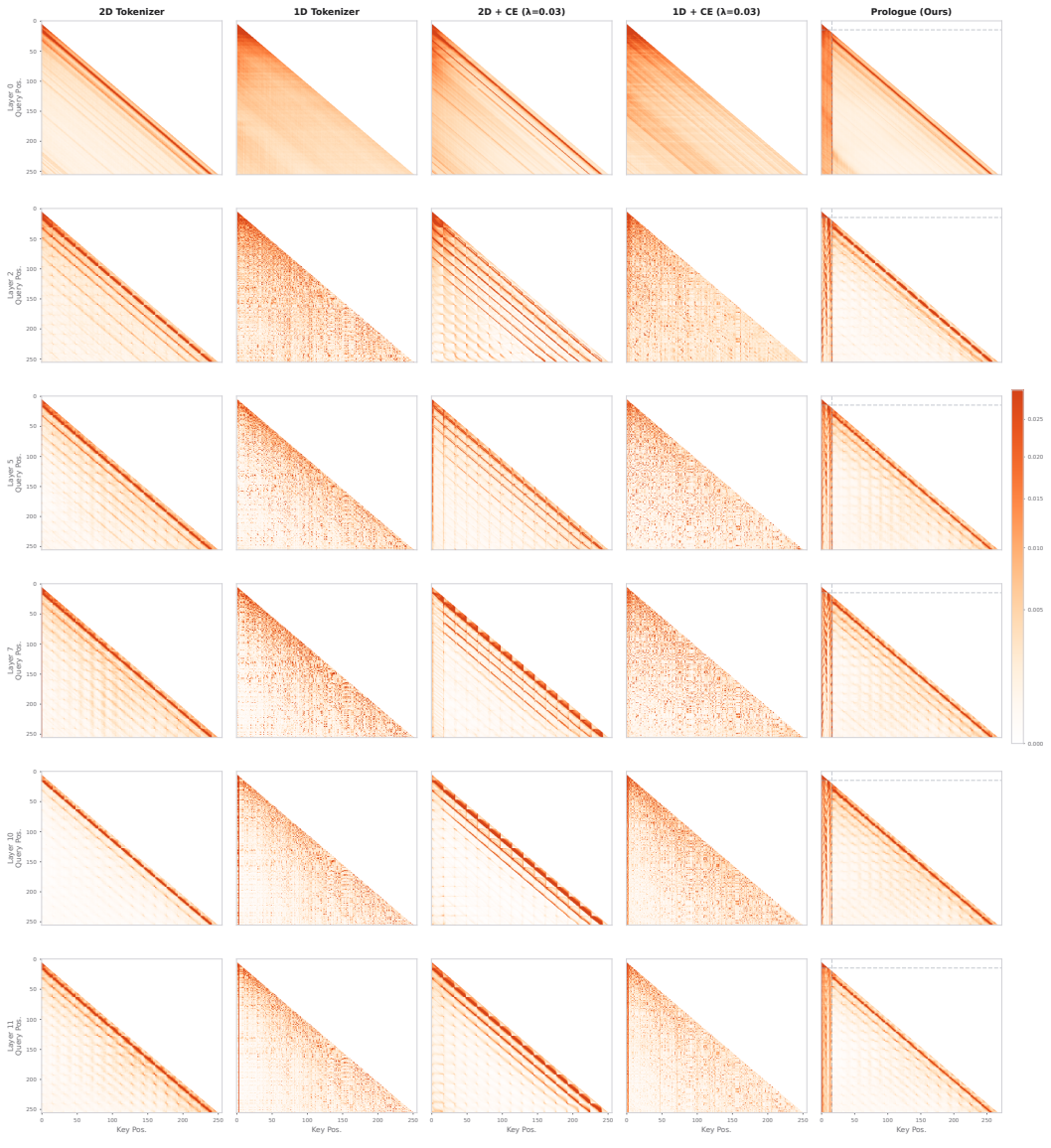


Figure A2: AR attention patterns (Layers 0/2/5/7/10/11; self-attention diagonal and first token masked). In shallow layers (0–2), all models show a uniform and local attention pattern, though the Prologue column is faintly visible. In middle layers (5–7), the Prologue column is most prominent, and pattern differences are most pronounced. In deep layers (10–11), the Prologue column persists; visual tokens attend to the prologue prefix across all depths.

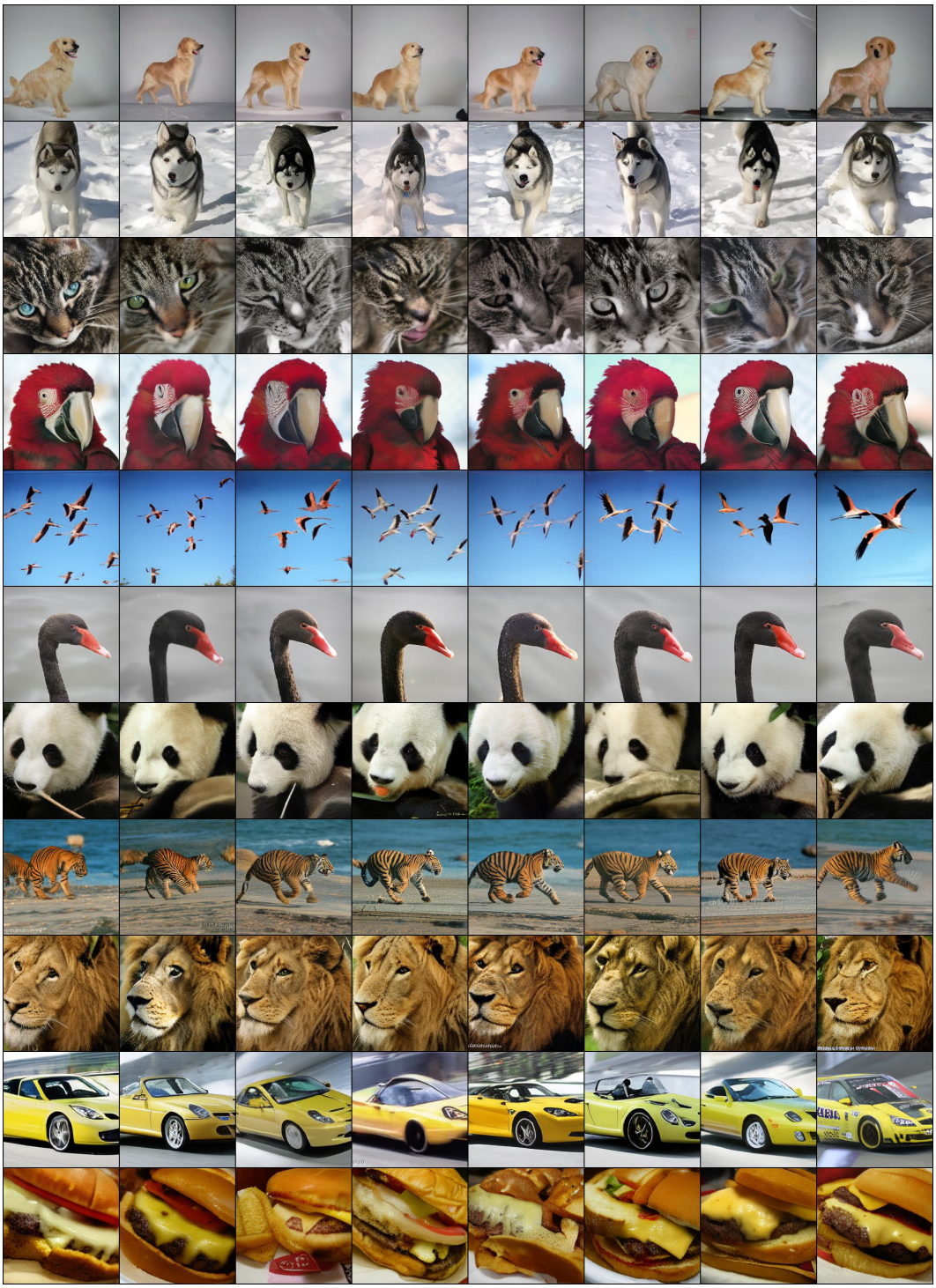


Figure A3: Prologue L–XL generation with **fixed prologue tokens** (11 classes, 8 visual resamples per class). Within each row, all 16 prologue tokens are held constant and only visual tokens are resampled. Samples share similar class identity and global layout while varying in texture and local details, suggesting that prologue tokens provide global semantic conditioning.

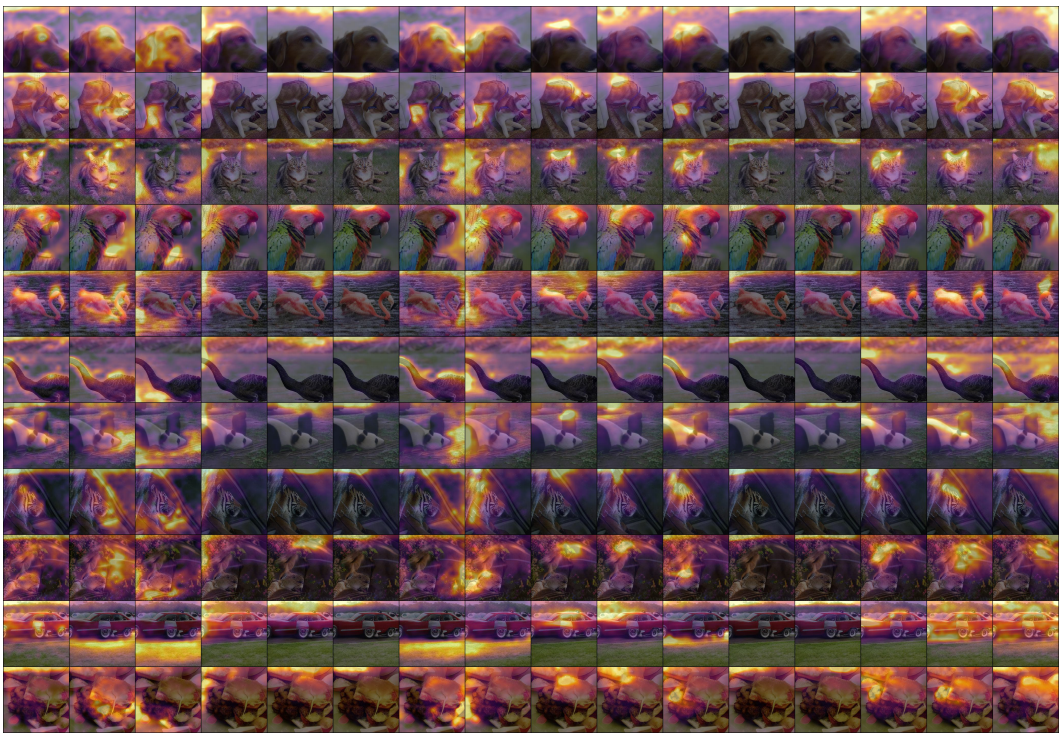


Figure A4: Attention heatmaps showing visual-token attention to each prologue token (averaged from the last 8 layers and heads). Prologue tokens provide rich semantic information to visual tokens about objects and background in different regions, indicating a learned spatial division of labor.

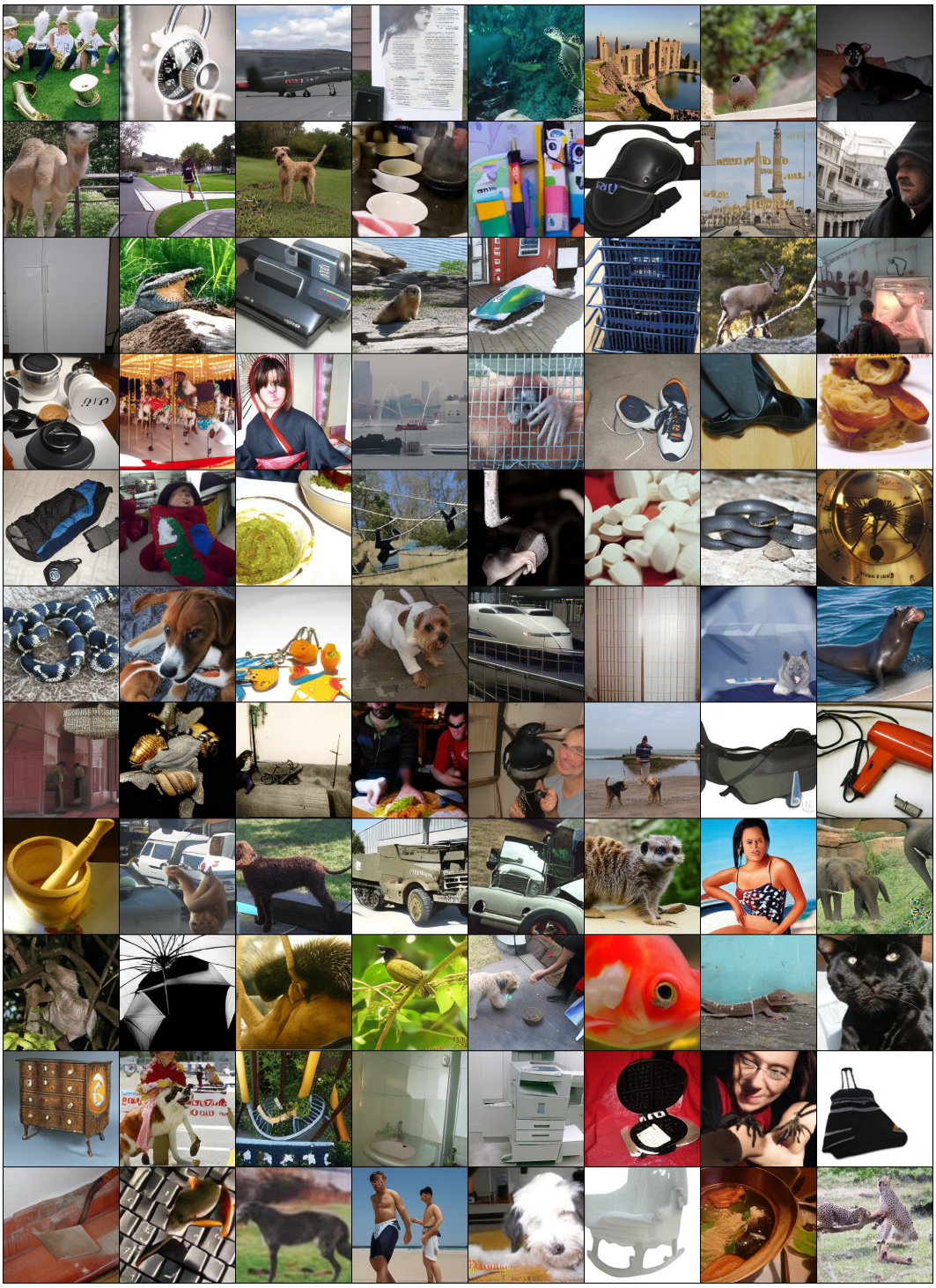


Figure A5: Uncurated samples from Prologue L–XL. Each grid contains 88 random ImageNet classes, with no repeats across the five grids. Sampling uses prologue CFG $= 0.7$ and visual CFG $\text{cosine}(\text{scale} = 2.25, p = 0.225)$. The model achieves gFID $= 1.46$ and IS $= 257.66$.

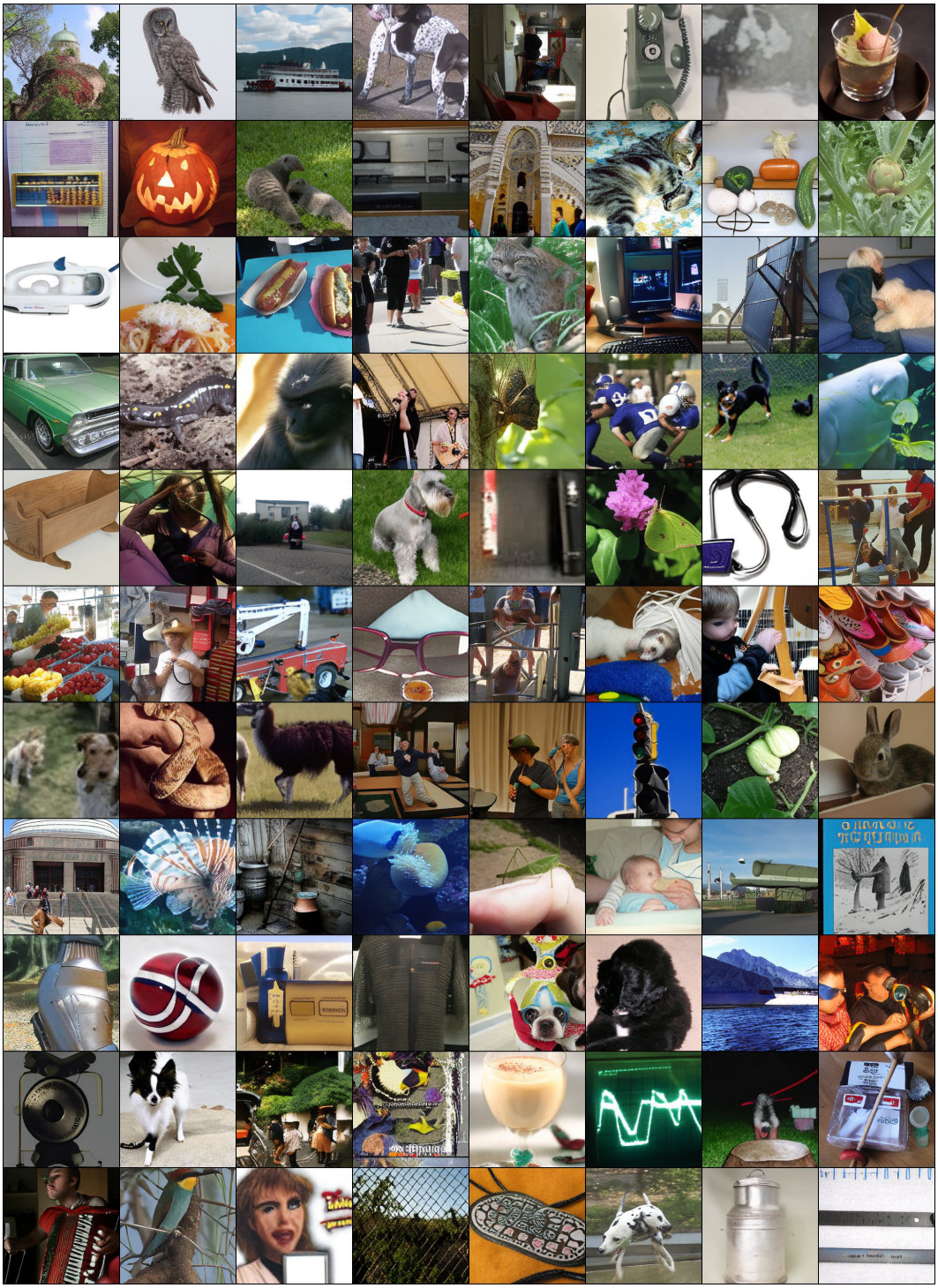


Figure A6: Uncurated samples from Prologue L–XL under the same setting as Fig. A5.

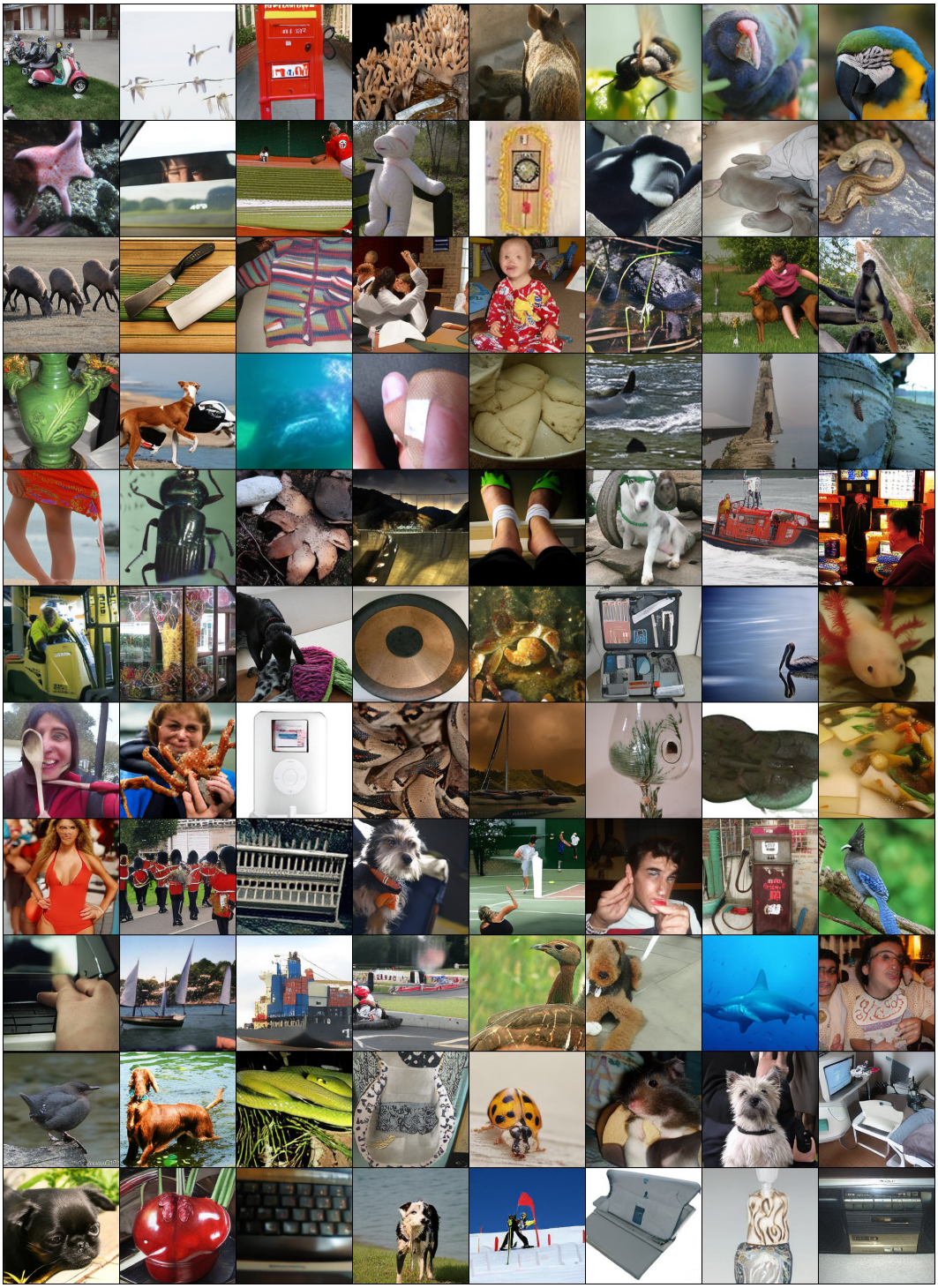


Figure A7: Uncurated samples from Prologue L–XL under the same setting as Fig. A5.

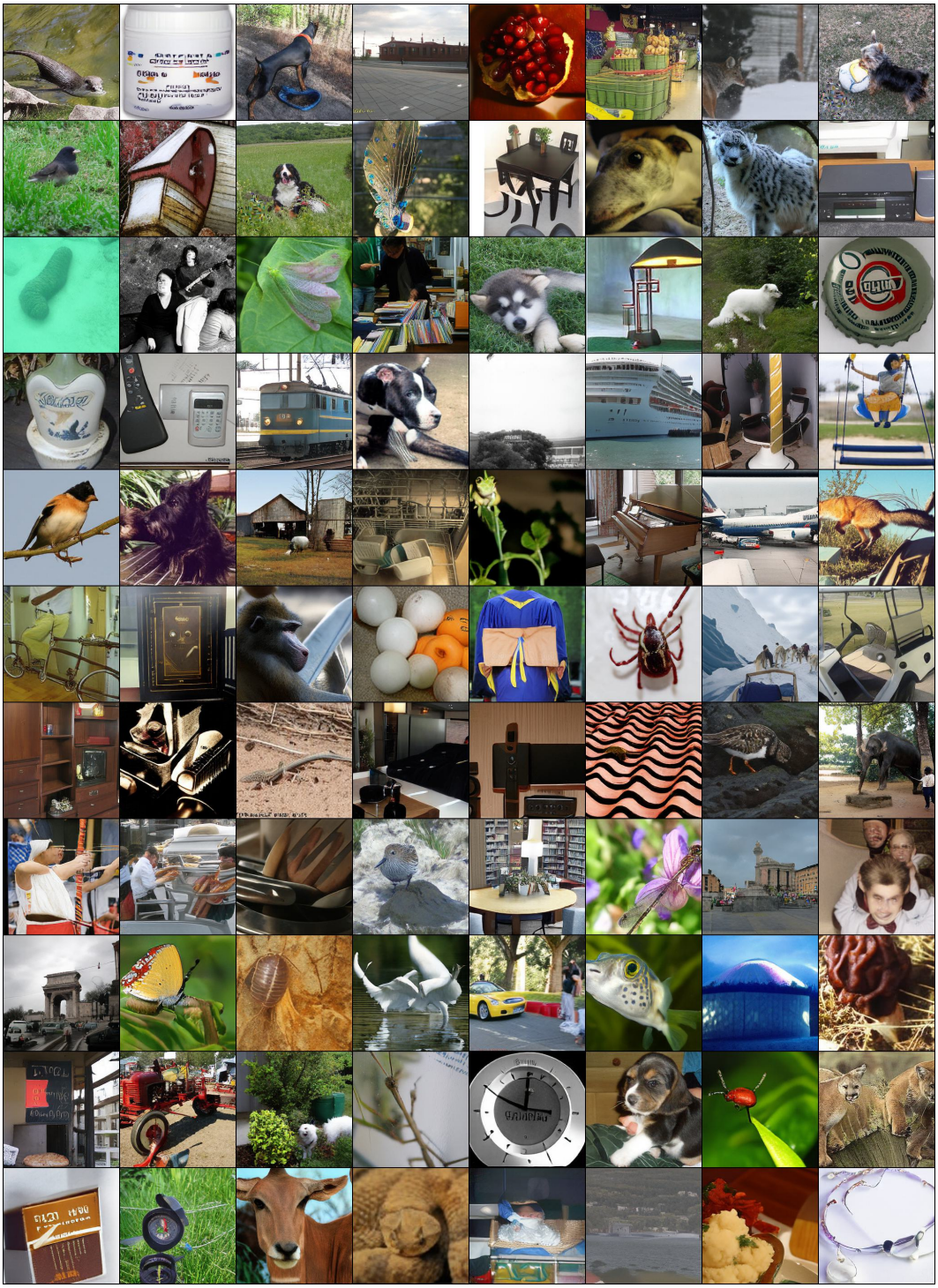


Figure A8: Uncurated samples from Prologue L–XL under the same setting as Fig. A5.

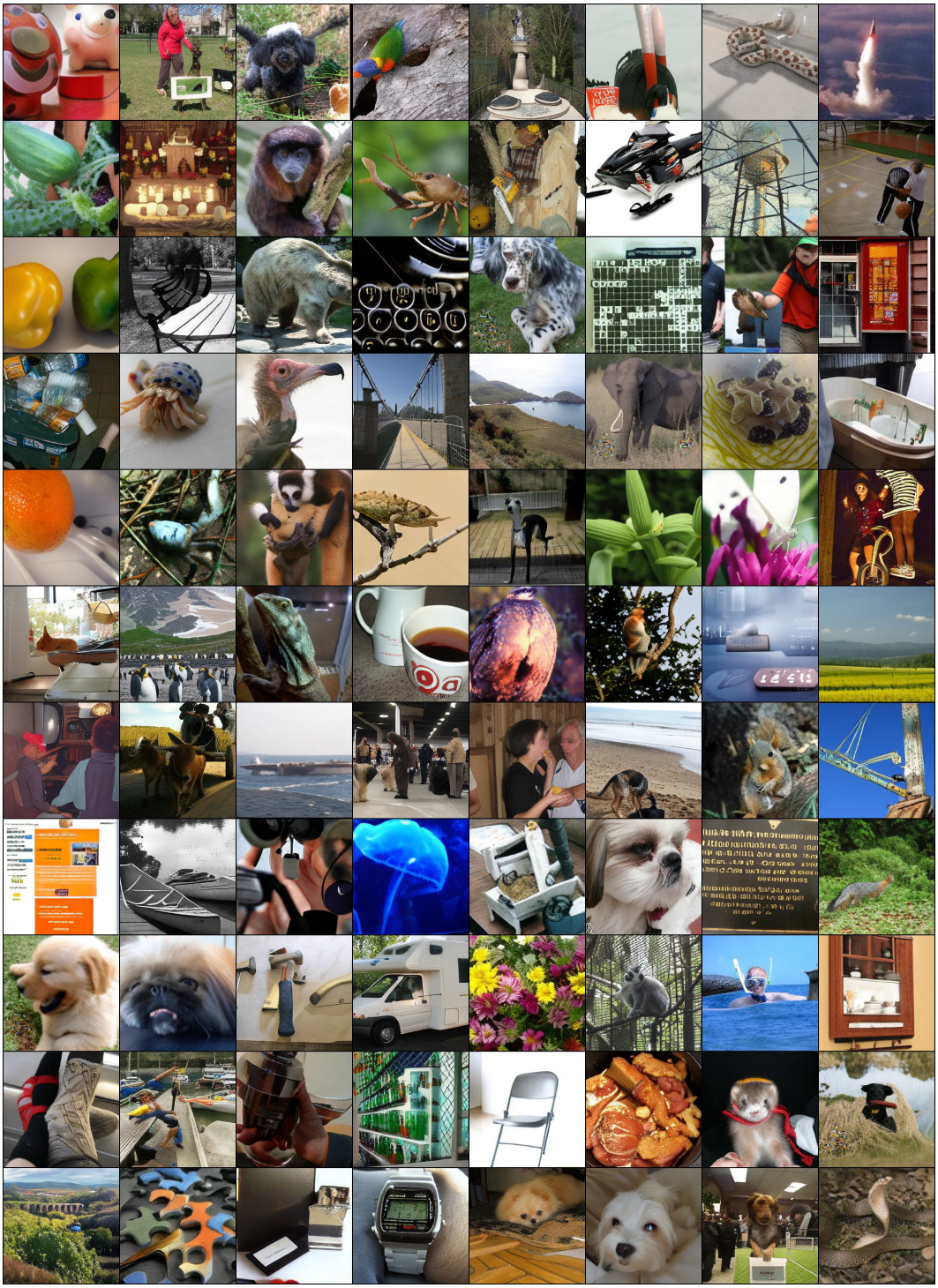


Figure A9: Uncurated samples from Prologue L–XL under the same setting as Fig. A5.



Figure A10: Uncurated samples from the 2D Tokenizer B–B model. Each grid contains 88 random ImageNet classes, with no repeats across the five grids. Sampling uses CFG $\mathrm{cosine}(\mathrm{scale} = 10, p = 1.5)$. The model achieves gFID $= 5.02$.

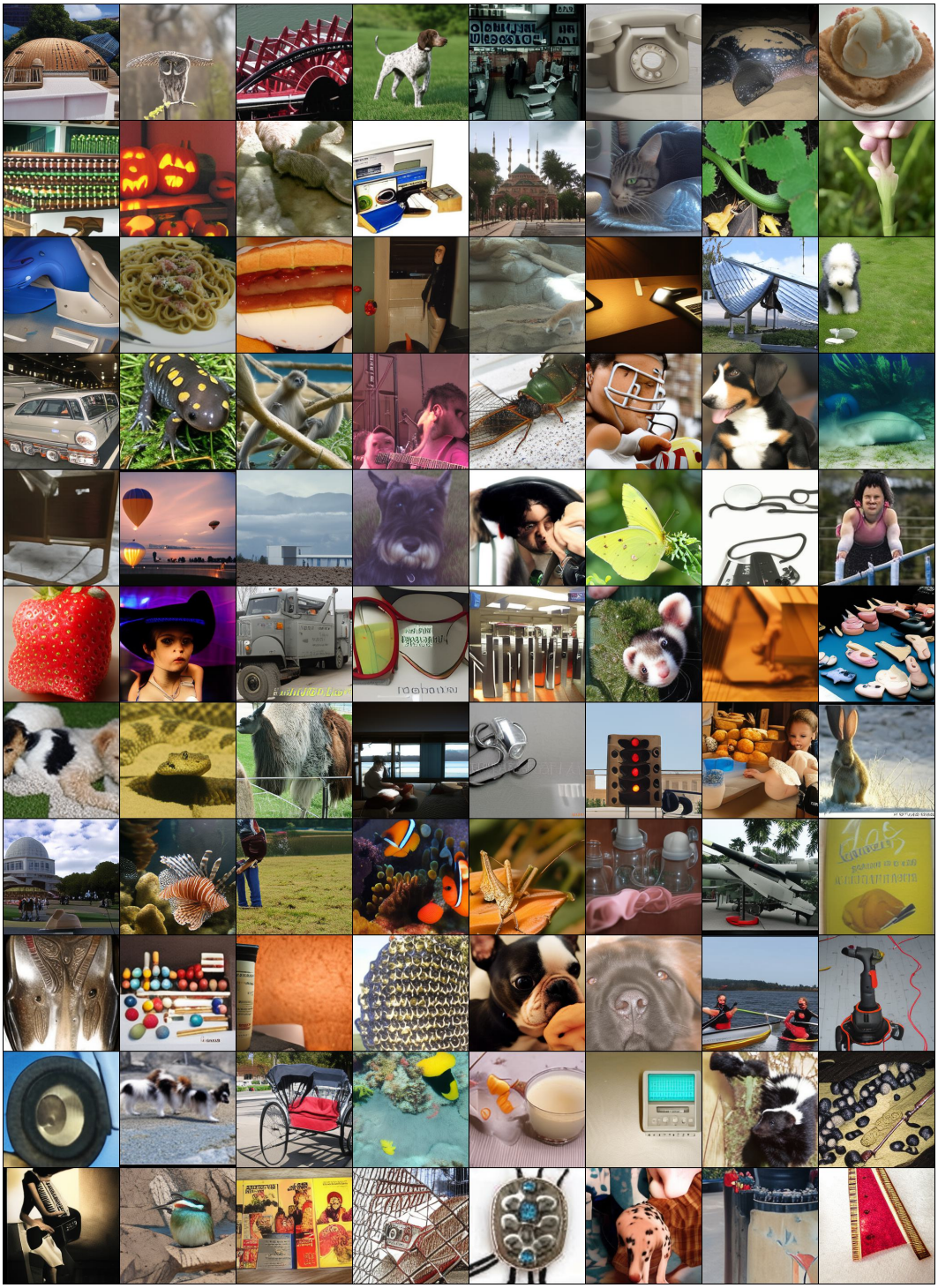


Figure A11: Uncurated samples from the 2D Tokenizer B–B model with the settings of Fig. A10.

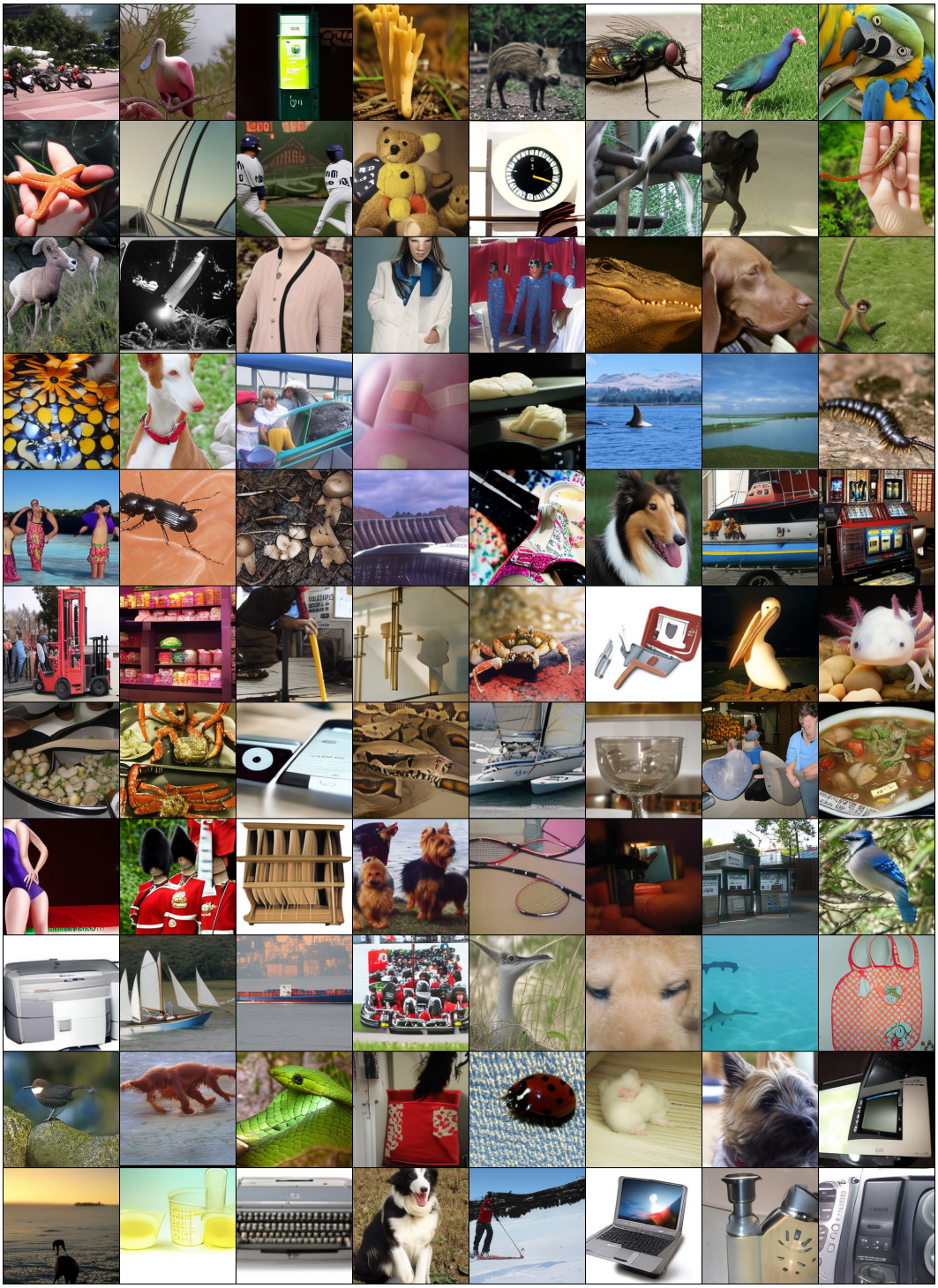


Figure A12: Uncurated samples from the 2D Tokenizer B–B model with the settings of Fig. A10.

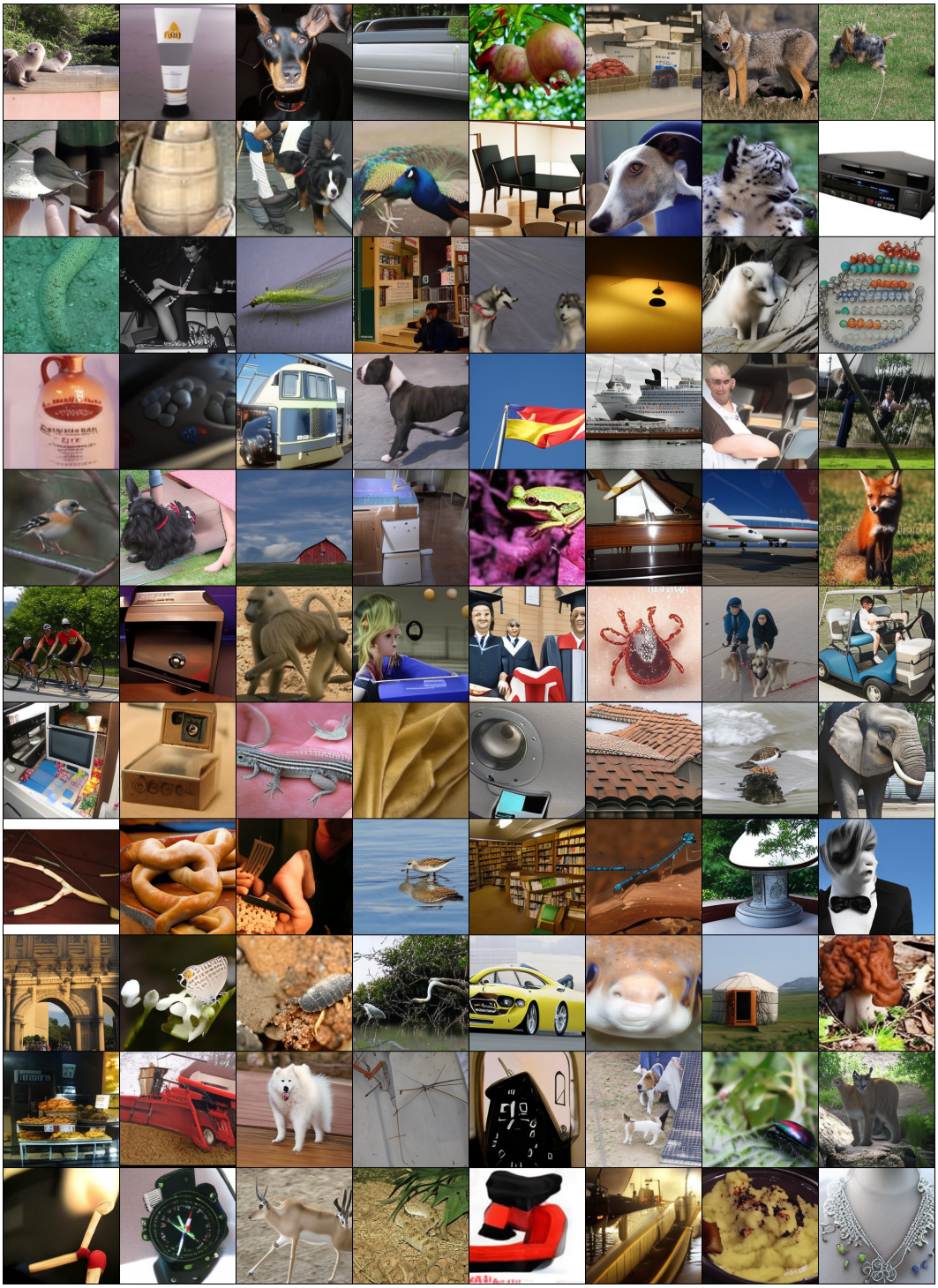


Figure A13: Uncurated samples from the 2D Tokenizer B–B model with the settings of Fig. A10.

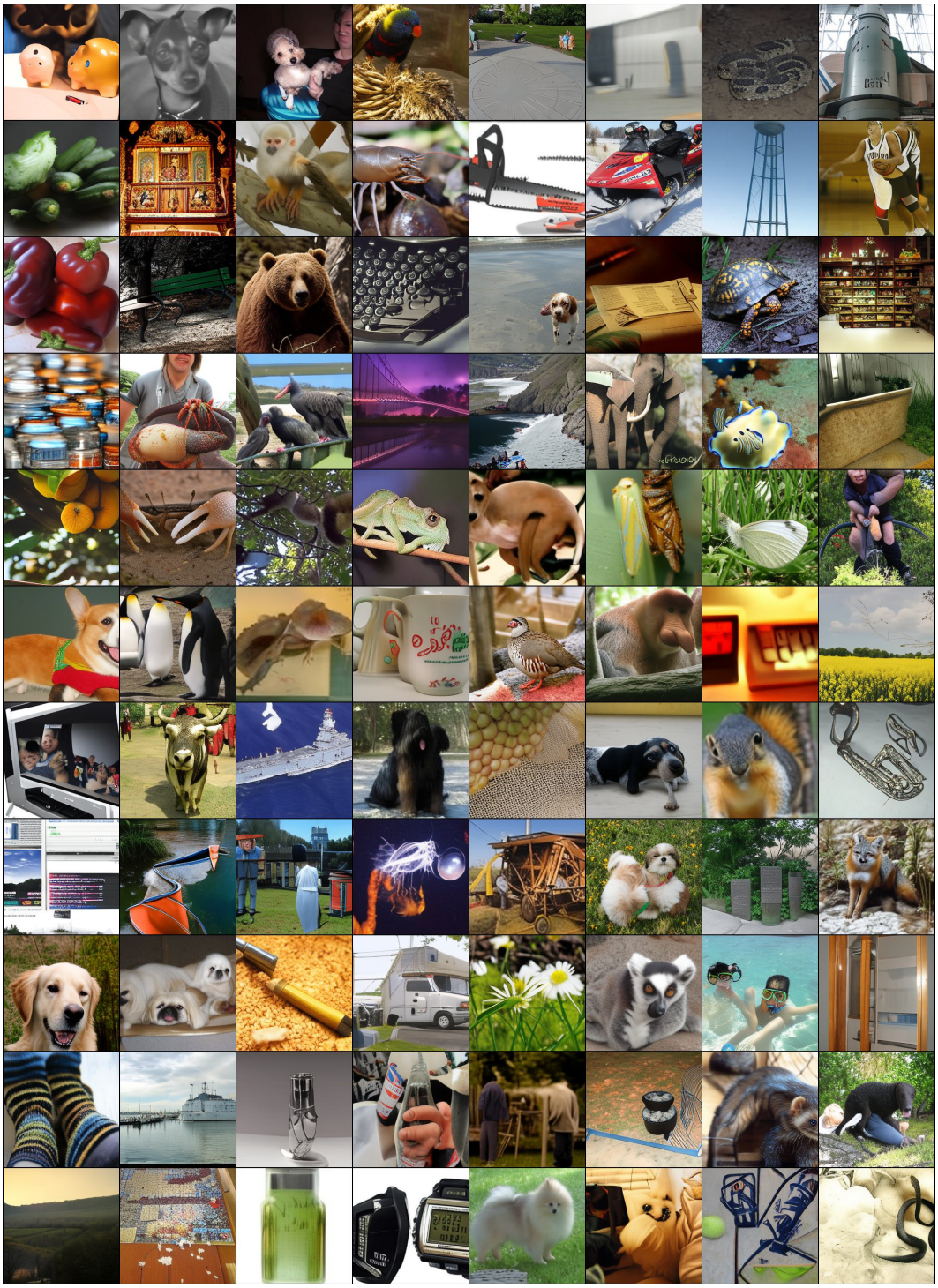


Figure A14: Uncurated samples from the 2D Tokenizer B–B model with the settings of Fig. A10.