Early Commitment Without Revision: Diagnosing Sampling Instability in Binary Masked Generative Models

Camilla Deckard
Independent
camilla.deckard2@gmail.com

Abstract

The unification of generative and discriminative objectives within a single visual backbone marks a major structural shift in modern computer vision. Masked binary generative models, such as BiGR, suggest that compact binary latent codes can jointly support image synthesis and classification. However, fine-grained generation remains inconsistent, particularly in high-frequency regions like faces and complex textures. This paper presents a systematic diagnostic analysis of generation instability in binary masked models. Through qualitative trajectory analysis, large-scale negative log-likelihood profiling across ImageNet classes, and controlled probability reversal measurement under deterministic sampling, we show that the iterative unmasking process is structurally compromised by early token commitment: approximately ***31% of Bernoulli probability estimates reverse direction*** *in early iterations, with reversal rates never dropping below 24%. We argue that this instability reflects a fundamental limitation of masked iterative generation that must be addressed to enable high-fidelity visual synthesis.*

1. Introduction

Generative modeling for computer vision is undergoing a paradigm transition. For over a decade, the dominant approach treated generation and recognition as fundamentally distinct problems, served by architecturally separate models. Diffusion models [6, 9, 10] achieved state-of-the-art synthesis quality but produced latent features poorly suited to discriminative downstream tasks—a consequence of what Balestriero and LeCun [1] identify as the core limitation of reconstruction-based learning: reconstruction fidelity does not imply perceptual separability. Self-supervised contrastive and masked image modeling methods [2, 5, 14] yielded strong representations but were not designed for generation.

A new generation of models is dismantling this divide. Masked autoregressive approaches—exemplified by MAGE [7] and BiGR [4]—demonstrate that a single model trained solely on generative reconstruction can achieve competitive performance on both axes simultaneously. BiGR is the first conditional generative model to unify both objectives, encoding images as sequences of binary latent codes and learning to predict masked tokens via a bidirectional Llama backbone, achieving FID-50k of 2.36 at 1.5B parameters and linear-probe accuracy of 69.8%—surpassing every prior conditional generative model by a substantial margin. Yet a gap persists: high-frequency content—faces, fingers, dense textures—remains systematically difficult to generate, and no mechanistic explanation has been offered. The present work fills that gap.

We argue that the source of this instability is structural, residing in BiGR's entropy-ordered sampling procedure. The model commits to binary token values iteratively, from most confident to least, with no mechanism to revisit earlier decisions. We show that this generates a compounding consistency problem: as the model conditions on an ever-growing set of unmasked tokens, its Bernoulli probability estimates for already-committed tokens shift substantially—and in 31% of early cases, reverse direction entirely. Nearly one in three early token decisions is actively contradicted by the model's own subsequent predictions. The implications extend beyond BiGR: any iterative masked generation strategy that (i) uses a fixed unmasking order, (ii) disallows token revision, and (iii) operates over a discrete latent space faces an analogous structural tension. Resolving it is essential to the maturation of unified generative-discriminative architectures as a general-purpose computer vision backbone.

2. Background: Binary Masked Generation

2.1. Binary Tokenization and the Bernoulli Latent Space

BiGR encodes an image $x \in \mathbb{R}^{3\times H\times W}$ into a sequence of binary latent codes $\{z^i\}_{i=1}^n$ using a Binary Autoencoder (B-AE) based on lookup-free quantization. Each token $z^i \in$

$\{0,1\}^K$ is obtained by thresholding the continuous encoder output:

$$z^i = \text{sign}(\zeta^i) = \mathbb{1}\{\zeta^i > 0\}, \tag{1}$$

where ζ^i is the pre-quantization feature at spatial position i, and K is the code dimension (16–32 bits in reported experiments). A vocabulary of 2^K token indices is thus implicitly defined. Crucially, BiGR operates directly on the binary code sequence rather than token indices, enabling a natural probabilistic interpretation via Bernoulli distributions.

2.2. Masked Modeling and the Binary Transcoder

Training proceeds by masking a fraction of tokens (cosine-scheduled) and learning to predict their binary values. The backbone is a bidirectional Llama transformer f_θ, which processes the partially-masked sequence and outputs continuous representations at masked positions. These are passed to a binary transcoder g_ϕ that implements a Bernoulli diffusion process [13], modeling the denoising distribution:

$$p_\phi\big(z^{t-1} \mid z^t\big) = \mathcal{B}\big(z^{t-1};\, S\big(g_\phi(z^t, t, h)\big)\big) \tag{2}$$

where S is the sigmoid function and h is the backbone feature. Training minimizes a weighted binary cross-entropy (wBCE) loss that accounts for bit-level imbalance between 0s and 1s:

$$\mathcal{L} = -\frac{1}{K}\sum_k w_k \left[y_k \log p_k + (1-y_k)\log(1-p_k)\right] \tag{3}$$

where y_k is the XOR target $(z^t \oplus z^0)[k]$, and p_k is the predicted Bernoulli probability for bit k. The weight w_k mitigates bit-class imbalance within each code.

2.3. Entropy-Ordered Sampling

At inference, BiGR generates images by iteratively unmasking tokens over N iterations (typically 20). The unmasking order is governed by a confidence score derived from the per-token binary entropy:

$$H = -\frac{1}{K}\sum_k \left[p_k \log_2 p_k + (1-p_k)\log_2(1-p_k)\right] \tag{4}$$

Confidence is computed as $1 - H$, with Gumbel noise added to promote diversity. At each iteration, the top fraction of masked tokens—by confidence—are sampled from their Bernoulli distributions and committed. The proportion of tokens unmasked per iteration follows a cosine schedule. Critically, once a token is committed, it remains fixed for all subsequent iterations; the process is strictly one-directional. This is the design choice whose consequences we investigate.

3. A Three-Step Diagnostic Analysis

Our investigation is motivated by a simple but underexplored question: does the confidence ordering of BiGR's sampling process produce consistent predictions over the course of generation? Specifically, do Bernoulli probability estimates for tokens committed in early iterations remain coherent with those estimated in later iterations, when the model has access to substantially more context? We answer this question through three progressively more rigorous steps. The model used throughout these experiments is BiGR-L-d24.

3.1. Step 1: Qualitative Evidence of Early-Stage Instability

We examined trajectories of Bernoulli probability estimates for committed tokens across generations of 8 diverse ImageNet samples. At each iteration, we tracked the model's updated probability estimates for tokens committed in prior steps. Token confidence was defined as $2|p - 0.5|$, the normalized distance from maximum uncertainty, consistent with BiGR's implementation.

Two representative cases illustrate the phenomenon (Figure 1). For a successfully generated chimpanzee, early-committed tokens retain high confidence across iterations: as unmasked tokens increase, their updated Bernoulli estimates remain concentrated in the high-confidence region, indicating stable global structure. For a distorted pufferfish, a contrasting pattern appears: a token committed with confidence above 0.9 in iteration 1 drops to the lower tail by iteration 2, indicating the model becomes less certain—or disagrees with—its earlier decision.

Instability is concentrated in early iterations: by around iteration 6 of 20, the gap between confidence distributions of committed and uncommitted tokens largely disappears. This supports a context-scarcity hypothesis: early predictions, made with limited anchor tokens, are more volatile than those under richer context. The pattern holds across all 8 samples, motivating large-scale quantification.

3.2. Step 2: Quantifying Instability via Negative Log-Likelihood

To verify this observation at scale, we ran the model over one image per class across all 1,000 ImageNet classes (1,000 total samples). At each iteration t, we computed the negative log-likelihood of previously-committed token values under the current Bernoulli distributions. For a committed token with binary value $z_i \in \{0,1\}^K$ and current probability estimate p_i, the per-token contribution to NLL is:

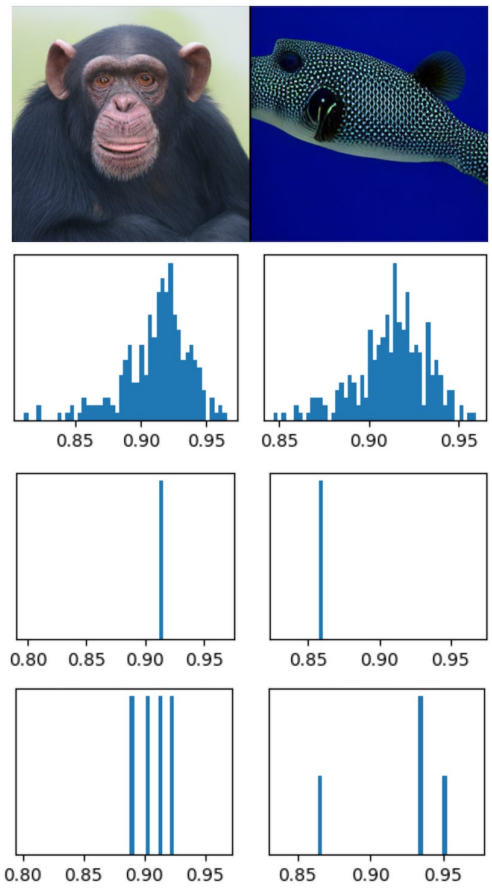

Figure 1. Top row: generated images. Second row: confidence distribution of tokens prior to commitment in iteration 1. Third and fourth rows: confidence distribution of committed tokens across iterations 2–3. For the chimpanzee, early-committed tokens remain highly confident. For the pufferfish, a token committed at iteration 1 drops to the lower tail by iteration 2, indicating early instability.

$$\mathrm{NLL}_i^{(t)} = -\frac{1}{K}\sum_k \Big[z_i[k] \log p_i^{(t)}[k] + (1 - z_i[k]) \log(1 - p_i^{(t)}[k]) \Big] \tag{5}$$

A high NLL at iteration t for tokens committed at iteration $s < t$ indicates that the model's current predictions are inconsistent with its own earlier commitments—operationally, that the generation process is self-contradictory. The mean NLL across all committed tokens is computed at each iteration and averaged across all 1,000 class samples.

The results (Figure 2) confirm the qualitative findings at scale. Mean NLL peaks at iteration 3 (approximately 0.96) and decays, stabilizing around 0.80–0.84 after iteration 10. The decay profile roughly aligns with the qualitative observation that instability attenuates near iteration 6: as the unmasked context grows, the model's predictions for committed tokens become progressively more consistent with their committed values. The NLL floor—rather than reaching the theoretical minimum—indicates persistent residual inconsistency throughout generation.

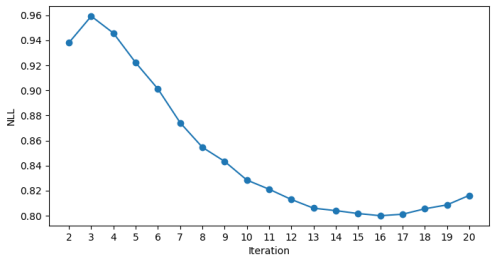

Figure 2. Mean NLL of committed tokens under current Bernoulli estimates, averaged across 1,000 ImageNet classes, as a function of generation iteration. NLL peaks at iteration 3 ($\approx$ 0.96) and decays to a stable floor around iteration 10 ($\approx$ 0.80–0.84).

3.3. Step 3: Isolating Probability Estimate Reversals

Elevated NLL admits two interpretations: (1) confidence reduction, where an estimate shifts *e.g.* from 0.9 to 0.7 but remains directionally correct, and (2) directional reversal, where an estimate shifts *e.g.* from 0.8 to 0.2, contradicting the committed value. The former reflects uncertainty; the latter indicates structural inconsistency—the model has committed to a value it later deems incorrect.

To isolate reversals, we removed stochastic sampling by using deterministic decoding: probabilities above 0.5 map to 1, below 0.5 to 0. Under this scheme, reversals cannot be attributed to sampling noise and must reflect genuine shifts in probability estimates. At each iteration t, we measured the percentage of values committed through iteration $t-1$ whose Bernoulli estimates crossed the 0.5 threshold relative to their commitment step. The experiment spans all 1,000 ImageNet classes.

The results in Fig. 3 are clear: about 31% of committed early values reverse direction shortly after commitment—nearly **one in three** early decisions is later **contradicted**. This rate persists, never dropping below 24%, indicating reversals are a structural feature of BiGR rather than a transient early-stage effect.

These findings refine the failure mode: instability is not just low confidence, but frequent self-contradiction in probability estimates. Without a revision mechanism, these contradictions propagate irreversibly into the final image, fixing corrupted global structure that later tokens must condition on and cannot correct.

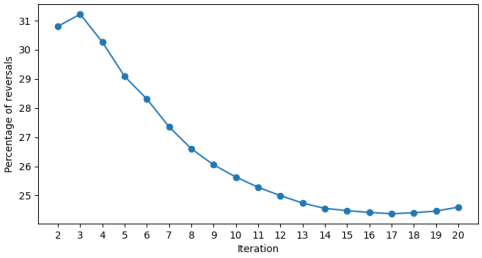

Figure 3. Percentage of committed binary values undergoing directional reversal of Bernoulli probability estimates, averaged across 1,000 ImageNet classes, as a function of generation iteration. Peak reversal rate: ∼31% at iteration 3. Floor: ∼24% at late iterations.

4. Implications For Masked Generative Modeling

The findings above establish a concrete failure mode in BiGR's generation strategy. We now situate this failure within the broader landscape of masked generative approaches and identify the directions it motivates.

4.1. A Structural Tension in Iterative Unmasking

The early-commitment problem is not specific to BiGR's hyperparameters or training setup. It arises from three shared design choices in masked generative models: (i) a fixed confidence-based unmasking order, (ii) no mechanism to revise committed tokens, and (iii) an effectively autoregressive information structure, where early tokens condition later ones but not vice versa—despite bidirectional attention during training. MaskGIT [3], MAGE [7], and related methods share this structure and are thus prone to similar dynamics, though the effect size depends on the latent space and probability parameterization.

The binary latent space may further amplify this issue relative to continuous alternatives. MAR [8] uses a diffusion-based continuous token distribution, allowing uncertainty to be represented more softly; a high-variance continuous commitment is less damaging than a discrete binary one that is later reversed. At the same time, the binary space offers advantages in compactness and separability—evidenced by BiGR's linear-probe performance—suggesting the solution lies in improved sampling strategies rather than abandoning binary representations.

4.2. Toward Revisable Generation

Two directions follow directly from this analysis. The first is iterative token revision: extending unmasking to allow previously committed tokens to be re-evaluated—and, if needed, re-sampled—based on updated probability estimates. This parallels re-masking strategies in diffusion schedulers and could be implemented as a lightweight pass over early tokens at each iteration. The trade-off is additional forward passes for the benefit of eliminating inconsistent commitments. However, our preliminary experiments along these lines did not yield improvements, suggesting that naive revision strategies may be insufficient.

The second is early-iteration conditioning augmentation: injecting extra structural signals (e.g., class embeddings, coarse layouts, or guidance from a low-resolution generation) during the initial iterations, where instability is highest and early tokens define global structure. This aligns with hierarchical approaches like VAR [12], which generate coarse structure before refinement, though adapting it to BiGR's masked framework would be required.

4.3. Relationship to Unified Generative-Discriminative Modeling

The broader significance of this analysis lies in what it reveals about the scalability of unified generative-discriminative architectures. BiGR's discriminative performance—69.8% linear-probe accuracy at 799M parameters, substantially outperforming LlamaGen (40.5%) [11], MAR-H (60.0%) [8], and prior generative methods—demonstrates that binary masked generation is a viable path toward general visual representations. The present work identifies the principal obstacle preventing this paradigm from also achieving the generation quality needed for high-fidelity visual synthesis: not the model capacity, not the tokenizer, but the sampling procedure.

Resolving early-commitment does not require retraining BiGR, as the instability arises at inference time from the model's Bernoulli distributions. Improved sampling strategies—guided by the NLL and reversal-rate metrics—can be applied directly to existing checkpoints. These metrics form a diagnostic toolkit for evaluating sampling quality in iterative masked generative models.

5. Conclusion

We analyze generation instability in BiGR, showing that entropy-ordered sampling is undermined by early token commitment. Across 1,000 ImageNet classes, 31% of Bernoulli estimates reverse after early-stage commitment, remaining above 24% throughout generation, identifying fine-grained failures as an inference-time issue rather than a limitation of binary latents or masked training.

These results highlight a key challenge for unified generative–discriminative vision models: achieving high-fidelity generation requires addressing commitment without revision. The proposed NLL and reversal metrics provide a foundation for revisable and conditioning-augmented sampling strategies.

References

[1] Randall Balestriero and Yann Lecun. How learning by reconstruction produces uninformative features for perception. In *Proceedings of the 41st International Conference on Machine Learning*, pages 2566–2585. PMLR, 2024. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 1

[3] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[4] Shaozhe Hao, Xuantong LIU, Xianbiao Qi, Shihao Zhao, Bojia Zi, Rong Xiao, Kai Han, and Kwan-Yee K. Wong. BiGR: Harnessing binary latent codes for image generation and improved visual representation capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022. 1

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 1

[7] Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2142–2152, 2023. 1, 4

[8] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems*, pages 56424–56445, 2024. 4

[9] William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2023. 1

[10] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 1

[11] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 4

[12] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: scalable image generation via next-scale prediction. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024. 4

[13] Ze Wang, Jiang Wang, Zicheng Liu, and Qiang Qiu. Binary latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22576–22585, 2023. 2

[14] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Loddon Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. 1