Annealed Relaxation of Speculative Decoding for Faster Autoregressive Image Generation

Xingyao Li[1], **Fengzhuo Zhang**[1], **Cunxiao Du**[2*], **Hui Ji**[1]
[1]National University of Singapore
[2]Sea AI Lab
xingyao@u.nus.edu, fzzhang@u.nus.edu, cnsdunm@gmail.com, matjh@nus.edu.sg

Abstract

Despite significant progress in auto-regressive image generation, inference remains slow due to the sequential nature of AR models and the ambiguity of image tokens, even when using speculative decoding. Recent works attempt to address this with relaxed speculative decoding but lack theoretical grounding. In this paper, we establish the theoretical basis of relaxed SD and propose COOL-SD, an annealed relaxation of speculative decoding built on two key insights. The first analyzes the total variation (TV) distance between the target model and relaxed speculative decoding and yields an optimal resampling distribution that minimizes an upper bound of the distance. The second uses perturbation analysis to reveal an annealing behaviour in relaxed speculative decoding, motivating our annealed design. Together, these insights enable COOL-SD to generate images faster with comparable quality, or achieve better quality at similar latency. Experiments validate the effectiveness of COOL-SD, showing consistent improvements over prior methods in speed-quality trade-offs.

Code — https://github.com/xingyaoL/COOL-SD

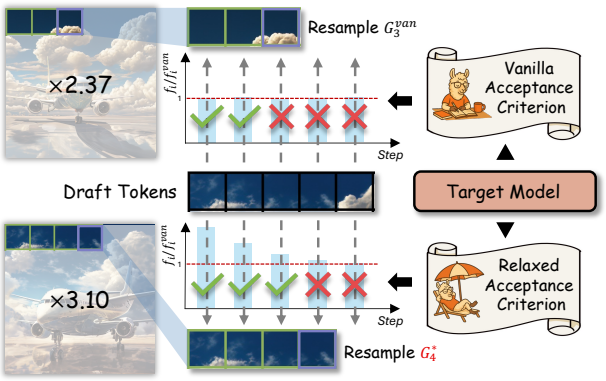


Figure 1: Illustration of vanilla SD and COOL-SD. By increasing the acceptance criterion f_i according to an annealing schedule, along with a principled resampling distribution G_i^*, we can further increase the inference speed.

1 Introduction

Auto-Regressive (AR) models have recently emerged as a powerful paradigm for image generation, often achieving performance comparable to or even surpassing that of diffusion-based methods (Sun et al. 2024a; Chen et al. 2025; Rombach et al. 2022; Esser et al. 2024). These models generate images in a sequential manner by predicting one token at a time, conditioned on all previously generated tokens. While this AR factorization is effective for modeling complex dependencies, it leads to substantial inference latency. Each token requires a separate forward pass, making the decoding process very costly for high-resolution images with thousands of tokens. Consequently, AR models are inefficient for real-time or interactive applications. Accelerating the decoding process is therefore a key challenge for making AR models more practical and widely applicable.

Speculative Decoding and Its Variants Speculative Decoding (SD) (Leviathan, Kalman, and Matias 2023; Chen et al. 2023) has emerged as an effective lossless approach for reducing the latency of auto-regressive generation, producing outputs from the exact target distribution while requiring fewer forward passes. SD leverages two models: a high-quality but slow target model and a smaller, faster draft model. The decoding process proceeds in two stages: drafting, where the draft model auto-regressively generates a sequence of tokens, and verification, where the target model evaluates the entire draft sequence in a single forward pass and accepts a prefix of tokens that match its predictions. If any token is rejected, the target model resamples to correct for distributional bias. The accepted tokens are appended to the prefix, and the process repeats. By reducing the number of calls to the target model, SD accelerates the process without altering the output distribution.

Despite its success in accelerating the Large Language Models (LLM) inference (Cai et al. 2024; Li et al. 2024b), SD's acceleration effect in image generation is limited. Prior studies attribute this to token selection ambiguity (Jang et al. 2024). That is, at many decoding steps, multiple candidate tokens receive nearly identical probabilities, increasing the difficulty for the draft model to accurately predict the target model's top-ranked tokens. This mismatch lowers the token acceptance rate during the verification stage, thereby reducing its overall efficiency. There have been several attempts to

*Corresponding author.

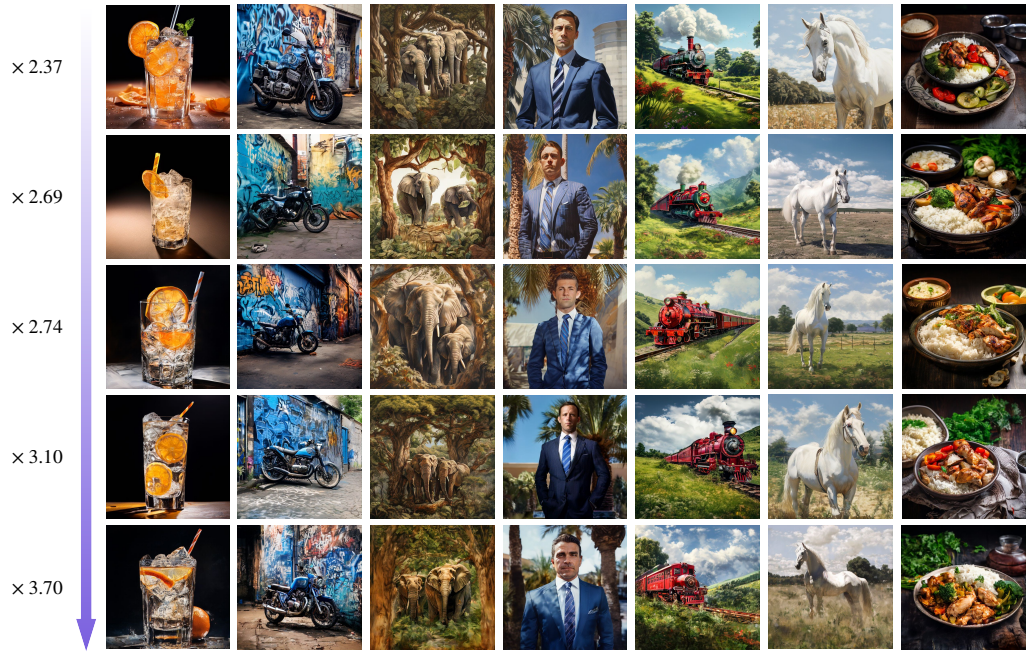


Figure 2: Qualitative results of COOL-SD. We demonstrate the trade-off between generation efficiency and image quality by comparing outputs from COOL-SD on Lumina-mGPT under different parameter settings. The speed-up factor is annotated to the left of each row. The first row shows images generated by Eagle-1 without any relaxation, serving as the baseline.

address this issue. LANTERN (Jang et al. 2024) introduces a relaxed verification scheme that permits limited divergence from the target distribution, while LANTERN++ (Park et al. 2025) improves efficiency via a static tree structure. However, both rely on heuristic, empirically driven designs and lack any mathematical account of their divergence from the target distribution. Moreover, their performance gains remain modest, with room for improvement in either image quality or generation speed. This calls for a more principled approach to relaxed SD, with mathematical analysis on the approximation to the target distribution, to achieve better quality at the same level of acceleration.

Main Idea and Contributions Driven by the need for a principled and effective relaxation of speculative decoding, this paper aims at developing an acceleration scheme that improves the speed-quality trade-off beyond vanilla SD, supported by rigorous analysis of distributional fidelity.

In this paper, we propose an acceleration technique that builds on a generalization of vanilla SD, replacing its rigid, exact-match acceptance test with a relaxed alternative. In our relaxed framework, each drafted token is accepted with a controllable probability f_i, rather than only when it exactly matches the target model's prediction. While this decouples the drafting and verification stages, thus enabling longer expected prefixes, it also introduces distributional drift. To address this, we derive token-wise resampling distributions $\{G_i^*\}_{i=1}^L$ by minimizing an almost-tight upper bound on the total variation (TV) distance between the output distribution and the target model. These optimal resampling distributions are specifically designed to closely approximate the target distribution under relaxed acceptance, and can be seamlessly integrated into any relaxed SD variant.

Furthermore, we showed that an *annealing property* of the acceptance criterion: for a fixed expected number of accepted tokens, the bound is minimized when acceptance probabilities decrease monotonically across positions. Motivated by this, we adopt an exponentially decaying acceptance schedule in relaxing f_i, and combine it with the our derived resampling distributions $\{G_i^*\}$. This leads to a more efficient and theoretically grounded decoding strategy, which we refer to as **COOL-SD**.

In short, we present a theoretically grounded and empirically effective framework for accelerating AR generation via a principled relaxation of SD. Our main contributions are:

- We revisit SD and study its relaxed framework that replaces the rigid exact-match acceptance test with general acceptance rules. Within this new framework, we present several insights for algorithm design of relaxed SD. Specifically, we derive the expected number of accepted tokens and a nearly-tight upper bound on the TV distance. Minimizing this bound yields token-wise optimal resampling distributions that closely approximate the target model. We further show that a monotonically decaying acceptance schedule minimizes the bound.
- Building on these insights, we propose a scheme which integrates the derived resampling distributions with an exponentially decaying acceptance schedule, namely COOL-SD. It consistently improves the speed-quality Pareto frontier over existing SD variants and offers a principled foundation for future advances in efficient AR image generation.

• Extensive experiments on image generation tasks validate the effectiveness of our approach, showing consistent improvements over existing related techniques in both speed and quality trade-offs. Applying the derived optimal resampling distributions to the existing relaxed speculative decoding method, LANTERN, also yields significant gains in the efficiency and fidelity trade-off.

2 Related Work

2.1 AR Image Generation

In addition to diffusion-based models (Rombach et al. 2022) and flow-matching methods (Esser et al. 2024), autoregressive generation of image tokens has recently emerged as a promising approach for unifying text and image generation (Sun et al. 2024a; Liu et al. 2024). LlamaGen (Sun et al. 2024a) is the first to adopt and fintune the LLaMA base model for image generation by training a VQ-VAE image tokenizer (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021). Building on this foundation, subsequent works have improved generation quality, modality coverage, and output diversity through better tokenizers, datasets, architectures, and training strategies (Liu et al. 2024; Chen et al. 2025; Wang et al. 2024; Team 2024; Xie et al. 2024). A notable advantage of these AR models is their ability to generate *multi-modal* tokens with a single model via a shared tokenizer (Liu et al. 2024). A comprehensive survey can be found in Xiong et al. (2024).

2.2 SD and Its Extensions

To mitigate the latency of AR generation in LLMs, Leviathan, Kalman, and Matias (2023) and Chen et al. (2023) propose Speculative Decoding (SD), which accelerates token generation by enabling a large model (the target model) to verify a sequence of draft tokens generated by a smaller model (the draft model). Follow-up works extend this linear-chain structure to tree-based variants for further acceleration (Miao et al. 2023; Cai et al. 2024; Du et al. 2024; Li et al. 2024b; Hu and Huang 2024). SpecTr (Sun et al. 2024c) formulates the problem from an optimal transport perspective and derives an algorithm that is optimal up to a multiplicative constant. Sun et al. (2024b) propose verifying all draft tokens as a single block, further improving the acceleration ratio. Yin et al. (2024) provide a theoretical analysis of SD, while Bachmann et al. (2025) and Zhong, Teku, and Tandon (2025) explore classifier-based verification to accept more draft tokens and reduce verification time. Although these works primarily target text generation, Jang et al. (2024) adapt SD to the image domain by relaxing the verification mechanism and propose LANTERN, significantly reducing generation latency. LANTERN++ (Park et al. 2025) further improves this approach by replacing the dynamic tree structure with a static one. However, the theoretical foundations of relaxed SD for image generation remain largely unexplored in the current literature. More discussions on related methods are provided in Appendix B.

Algorithm 1: **Vanilla SD** $(-)$ / **COOL-SD** $(+)$

Input: target model P, draft model Q, prefix $\texttt{pt}$, draft length L.

Procedure:

1: *// Drafting Stage:*
2: Generate L draft tokens $\tilde{x}_{1:L}$ via $\tilde{x}_i \sim Q(\cdot \,|\, \texttt{pt}, \tilde{x}_{1:i-1})$ for $i \in [L]$.
3: *// Verification Stage:*
4: Set $\tau = 0$, and calculate the values of $P(\tilde{x}_i \,|\, \texttt{pt}, \tilde{x}_{1:i-1})$ for $i \in [L]$ in parallel.
5: **for** $i = 1, \ldots, L$ **do**
6: Sample $r_i \sim \text{Uniform}([0, 1])$
7: **if** $r_i \leq f_i^{\text{van}}(\tilde{x}_{1:i}, P, Q)$ in Eqn. (1) **then** $-$
8: **if** $r_i \leq f_i(\tilde{x}_{1:i}, P, Q; \omega_i)$ in Eqn. (7) & (10) **then** $+$
9: Set $\tau = i$ and $x_i = \tilde{x}_i$.
10: **else**
11: Sample $x_i \sim G_i^{\text{van}}(\tilde{x}_{1:i-1}, P, Q)$ in Eqn. (2) $-$
12: Sample $x_i \sim G_i^{*}(\tilde{x}_{1:i-1}, P, Q)$ in Eqn. (6) $+$
13: **break.**
14: **end if**
15: **end for**
16: **if** $\tau = L$ **then** sample $x_{L+1} \sim P(\cdot \,|\, \texttt{pt}, x_{1:L})$ **end if**
17: Return $x_{1:\tau+1}$.

3 Method

3.1 Speculative Decoding

SD (Chen et al. 2023) is an acceleration technique for AR generation that enables the simultaneous generation of multiple tokens. We present the vanilla SD algorithm (Chen et al. 2023; Leviathan, Kalman, and Matias 2023) in Algorithm 1. It involves two models: a target model P, and a faster draft model Q. In the context of AR image generation, the target model P is typically a large image generation model such as LlamaGen (Sun et al. 2024a) or LuminamGPT (Liu et al. 2024). The draft model Q is a significantly smaller neural network (thus faster), and is trained to predict the behavior of the target model P. Given the prefix $\texttt{pt}$ to decode, which includes the input prompt and the previously generated tokens, the target (resp. draft) model predicts the next token by sampling from $P(\cdot \,|\, \texttt{pt})$ (resp. $Q(\cdot \,|\, \texttt{pt})$). One round of SD consists of two main stages: the *drafting* stage and the *verification* stage.

1. **Drafting Stage**: In this stage, SD uses the draft model Q to generate subsequent L candidate tokens $\tilde{x}_i \sim Q(\cdot \,|\, \texttt{pt}, \tilde{x}_{1:i-1})$ for $i \in [L]$ in an AR manner, shown in Line 2 of Algorithm 1.

2. **Verification Stage**: The target model P first computes the corresponding token distributions $P(\cdot \,|\, \texttt{pt}, \tilde{x}_{1:i-1})$ in parallel for $i \in [L+1]$ (Line 4). After that, each drafted token $\tilde{x}_i$ is independently accepted with probability (Line 7)

$$f_i^{\text{van}}(\tilde{x}_{1:i}, P, Q) := \min\left\{1, \frac{P(\tilde{x}_i \,|\, \texttt{pt}, \tilde{x}_{1:i-1})}{Q(\tilde{x}_i \,|\, \texttt{pt}, \tilde{x}_{1:i-1})}\right\}. \quad (1)$$

Let τ denote the number of accepted tokens out of the L

drafted candidates. If all drafted tokens are accepted, i.e. $\tau = L$, as the target model has already computed the sampling probability $P(\cdot \mid \mathtt{pt}, x_{1:L})$ for x_{L+1} during the drafting stage, we can obtain a total of $L+1$ tokens in this SD round (Line 16). If a token is rejected at some position $\tau+1$, the distribution will be corrected by sampling the $\tau+1$-th token from the resampling distribution $G^{\text{van}}_{\tau+1}(\tilde{x}_{1:\tau}, P, Q)$ (Line 11). $G^{\text{van}}_{\tau+1}(\tilde{x}_{1:\tau}, P, Q)$ is carefully designed to ensure that the generated tokens $\tilde{x}_{1:\tau+1}$ follow the target model's distribution, defined as

$$G^{\text{van}}_{\tau+1}\big(\cdot \mid (\mathtt{pt}, \tilde{x}_{1:\tau}), P, Q\big) = \mathrm{Norm}\Big(\big[P(\cdot \mid \mathtt{pt}, \tilde{x}_{1:\tau}) - Q(\cdot \mid \mathtt{pt}, \tilde{x}_{1:\tau})\big]_+\Big). \quad (2)$$

Here $\mathrm{Norm}(\cdot)$ normalizes the input to a probability distribution and $[\,\cdot\,]_+$ is defined as $\max\{0, \cdot\}$. The choice of $G^{\text{van}}_{\tau+1}$ effectively rectifies the bias introduced by the draft model, as presented in Theorem 1.

Theorem 1 (Speculative Decoding Recovers Unbiased Target Distribution (Chen et al. 2023))**.** *The output tokens $x_{1:\tau+1}$ obtained with vanilla SD follows the target model distribution $x_{1:\tau+1} \sim P(\cdot \mid \mathtt{pt})$.*

While vanilla SD is originally proposed for accelerating LLMs, Jang et al. (2024) discovered its inefficiency in AR image generation and enhance inference efficiency by increasing the acceptance probability. LANTERN (Jang et al. 2024) leverages the latent proximity of image tokens by aggregating the probabilities of a candidate token's k-nearest neighbors into the candidate's own probability mass. Following this, LANTERN++ (Park et al. 2025) replaces dynamic tree-based SD with a static tree structure, which demonstrates improved performance on image generation tasks, and refines the design of relaxation parameters. Specifically, LANTERN++ (Park et al. 2025) modify the candidate token's probability distribution to

$$P^{k,\lambda}(x \mid \mathtt{pt}, \tilde{x}_{1:i-1}) := \begin{cases} \sum_{x \in A^{k,\lambda}(\tilde{x}_i)} P(x \mid \mathtt{pt}, \tilde{x}_{1:i-1}), & \text{if } x = \tilde{x}_i, \\ 0, & \text{if } x \in A^{k,\lambda}(\tilde{x}_i) \text{ and } x \neq \tilde{x}_i, \\ P(x \mid \mathtt{pt}, \tilde{x}_{1:i-1}), & \text{otherwise.} \end{cases}$$

Here, $A^{k,\lambda}(\tilde{x}_i)$ denotes a selected subset in the k-nearest neighbors set of $\tilde{x}_i$ that satisfy $\sum_{x \in A^{k,\lambda}(\tilde{x}) \setminus \tilde{x}} P(x \mid \mathtt{pt}, \tilde{x}_{1:i-1}) < \lambda P(\tilde{x} \mid \mathtt{pt}, \tilde{x}_{1:i-1})$.

Accordingly, the acceptance criterion in Line 7 of Algorithm 1 is redefined as

$$f^{k,\lambda}_i(\tilde{x}_{1:i}, P, Q) := \min\left\{1, \frac{\sum_{x \in A_{k,\lambda}(\tilde{x}_i)} P(x \mid \mathtt{pt}, \tilde{x}_{1:i-1})}{Q(\tilde{x}_i \mid \mathtt{pt}, \tilde{x}_{1:i-1})}\right\},$$

and the corresponding resampling distribution in Line 11 is modified to be

$$G^{k,\lambda}_{\tau+1}\big(\cdot \mid (\mathtt{pt}, \tilde{x}_{1:\tau}), P, Q\big) := \mathrm{Norm}\Big(\big[P^{k,\lambda}(\cdot \mid \mathtt{pt}, \tilde{x}_{1:\tau}) - Q(\cdot \mid \mathtt{pt}, \tilde{x}_{1:\tau})\big]_+\Big).$$

However, we argue that the choice of $\{f^{k,\lambda}_i\}_{i=1}^L$ and $\{G^{k,\lambda}_i\}_{i=1}^L$ lacks theoretical justification. Consequently, the fidelity of the images generated by LANTERN remains uncertain. In the following, we conduct theoretical analysis of relaxed SD in its general form.

3.2 Annealed Relaxation of SD

A key limitation in improving the acceptance rate of vanilla SD lies in the rigidity of its acceptance criterion, as defined in Eqn. (1). Relaxing this criterion can increase the likelihood of accepting drafted tokens, particularly beneficial for image generation due to the semantic ambiguity of image tokens (Jang et al. 2024). However, the acceptance functions f^{van}_i and resampling distributions G^{van}_i proposed in Chen et al. (2023) were shown to be optimal solutions to a maximal coupling problem (Sun et al. 2024c). Consequently, any relaxation of the acceptance criterion inevitably introduces bias that cannot be perfectly corrected by adjusting the resampling distributions.

Relaxed Acceptance Criteria and Resampling Distribution Motivated by the aforementioned unavoidable bias, we analyze the trade-off between efficiency and fidelity loss in any relaxed SD algorithm. To encompass a broad class of potential algorithms, we generalize the acceptance criteria (Line 7) and the corresponding resampling distributions (Line 11) in Algorithm 1 as $\{f_i\}_{i=1}^L$ and $\{G_i\}_{i=1}^L$, respectively. We require only that each f_i takes $\tilde{x}_{1:i}$, P, and G, and each G_i takes $\tilde{x}_{1:i-1}$, P, and G as inputs. We refer to this general framework as *Relaxed* SD.

To enable a principled trade-off between efficiency and fidelity, we begin by formally quantifying both concepts. We measure *efficiency* by the expected number of accepted tokens, denoted as τ. The *fidelity loss* is evaluated using the total variation (TV) distance, a standard metric for quantifying the discrepancy between probability distributions. Specifically, the TV distance between two distributions P_1 and P_2 over a discrete support $\mathcal{X}$ is defined as

$$\mathrm{TV}(P_1, P_2) := \sum_{x \in \mathcal{X}} |P_1(x) - P_2(x)|/2.$$

The length of the generated token sequence $x_{1:\tau+1}$ is a random variable, influenced by the specific choices of the acceptance criteria $\{f_i\}_{i=1}^L$. For instance, a design with lower acceptance probabilities tends to yield shorter sequences (i.e., smaller τ). Therefore, to enable a fair comparison of the fidelity loss across different designs, we *virtually* extend each generated sequence to a fixed length of $L+1$ by sampling the remaining tokens $x_{\tau+2:L+1}$ from the target model, i.e., $x_i \sim P(\cdot \mid \mathtt{pt}, x_{1:i-1})$ for $i \in \{\tau+2, \ldots, L+1\}$. Since these additional tokens are generated from the target model P, this virtual extension does not introduce any extra fidelity loss. We denote the resulting full sequence distribution as $\hat{P}$, i.e., $x_{1:L+1} \sim \hat{P}(\cdot \mid \mathtt{pt}) := \hat{P}_{X_{1:L+1}}$. For notational consistency, we denote the target distribution as $P(\cdot \mid, \mathtt{pt}) := P_{X_{1:L+1}}$.

We state our analysis in Theorem 2. For simplicity, we omit the notation for the prefixed tokens $\mathtt{pt}$ in f_i, G_i, P, and Q throughout the remainder of the manuscript.

Theorem 2. *For the tokens generated by Relaxed SD with acceptance criteria $\{f_i\}_{i=1}^L$ and resampling distributions $\{G_i\}_{i=1}^L$ in Algorithm 1, the expectation of the number of accepted tokens is*

$$\mathbb{E}[\tau+1] = 1 + \sum_{i=1}^{L} \sum_{\tilde{x}_{1:i} \in \mathcal{X}^i} Q(\tilde{x}_{1:i}) \prod_{j=1}^{i} f_j(\tilde{x}_{1:j}, P, Q). \quad (3)$$

The total variation between $\hat{P}$ and P is upper bounded as

$$\begin{aligned} &\mathrm{TV}(\hat{P}_{X_{1:L+1}}, P_{X_{1:L+1}}) \quad (4)\\ &\leq \frac{1}{2} \sum_{i=0}^{L-1} \sum_{x_{1:i+1} \in \mathcal{X}^{i+1}} Q(x_{1:i}) \prod_{k=1}^{i} f_k(x_{1:k}, P, Q) \cdot \\ &\quad \Big| Q(x_{i+1} \,|\, x_{1:i}) \cdot f_{i+1}(x_{1:i+1}, P, Q) - P(x_{i+1} \,|\, x_{1:i}) \\ &\quad + G_{i+1}(x_{i+1} \,|\, x_{1:i}, P, Q) \\ &\quad \cdot \sum_{\tilde{x}_{i+1} \in \mathcal{X}} \Big[1 - f_{i+1}\big((x_{1:i}, \tilde{x}_{i+1}), P, Q\big)\Big] Q(\tilde{x}_{i+1} \,|\, x_{1:i}) \Big|. \end{aligned}$$

Given $\{f_i\}_{i=1}^L$, one minimizer $\{G_i\}_{i=1}^L$ of this upper bound is $G_{i+1}^*(\cdot \mid \tilde{x}_{1:i}, P, Q) =$

$$\mathrm{Norm}\Big(\big[P(\cdot|\tilde{x}_{1:i}) - Q(\cdot|\tilde{x}_{1:i}) f_{i+1}\big((\tilde{x}_{1:i}, \cdot), P, Q\big)\big]_+\Big). \quad (5)$$

The proof of Theorem 2 is provided in Appendix D. The theorem first derives the expected accepted sequence length under P and Q in Eqn. (3), then establishes an upper bound on the total variation between the relaxed resulting distribution $\hat{P}$ and the target distribution P. This upper bound is almost-tight in the sense that when f_i and G_i are chosen as f_i^{van} and G_i^{van}, the bound reduces to 0. Finally, we derive the optimal resampling distribution G_i^* that minimizes this bound given any acceptance criteria $\{f_i\}_{i=1}^L$, which coincides with G_i^{van} under the vanilla SD setup with f_i^{van}.

Design of G_i. We further characterize the property of the minimizer $\{G_i^*\}_{i=1}^L$ in the following proposition.

Proposition 1. *If $f_i(\cdot \mid x_{1:i}, P, Q) \geq f_i^{\mathrm{van}}(\cdot \mid x_{1:i}, P, Q)$ almost surely for all $x_{1:i} \in \mathcal{X}^i$ and $i \in [L]$, the following holds almost surely*

$$G_i^*(\cdot \mid x_{1:i-1}, P, Q) = G_i^{\mathrm{van}}(\cdot \mid x_{1:i-1}, P, Q). \quad (6)$$

The proof is provided in Appendix E. This result reveals a new property of $G_i^{\mathrm{van}}(\cdot \mid x_{1:i-1}, P, Q)$ that it minimizes the TV upper bound when $f_i(\cdot \mid x_{1:i}, P, Q) \geq f_i^{\mathrm{van}}(\cdot \mid x_{1:i}, P, Q)$. This also provides a simple and explicit expression of the resampling distribution when the acceptance probability is unchanged, i.e., identical to that in vanilla SD. Experimental results corroborating this observation are presented in Section 5.

Design of f_i. For clarity and conciseness, we now specify a concrete parameterization of the acceptance functions. We emphasize that the annealing property we present still holds for general choices of f_i, as formally shown in Appendix F. Since the acceptance threshold functions $f_i(\tilde{x}_{1:i}, P, Q)_{i=1}^L$ can take various forms, we adopt a simple yet effective instantiation by introducing a multiplicative relaxation parameter ω_i into their formulation:

$$f_i(\tilde{x}_{1:i}, P, Q; \omega_i) = \min\left\{1, \frac{\omega_i \cdot P(\tilde{x}_i \,|\, \mathtt{pt}, \tilde{x}_{1:i-1})}{Q(\tilde{x}_i \,|\, \mathtt{pt}, \tilde{x}_{1:i-1})}\right\}. \quad (7)$$

A straightforward choice is to set

$$\text{(UNIFORMRSD)} \qquad \omega_i = \delta \text{ for all } i \in [L],$$

which we refer to as Uniform Relaxation of Speculative Decoding (UNIFORMRSD) in the following, and we refer to δ as the relaxation budget.

To further improve the efficiency of relaxed SD, we conduct a perturbation analysis on the acceptance criterion f_i, to shed light on our choices of parameter ω_i other than a fixed schedule. In the following, we consider the perturbed $\{f_i\}_{i=1}^L$ when $L = 2$ as

$$\tilde{f}_1(x_1, c_1, P, Q; \omega_1) = f_1(x_1, P, Q; \omega_1) + c_1,$$

$$\tilde{f}_2((x_1, x_2), c_2, P, Q; \omega_2) = f_2((x_1, x_2), P, Q; \omega_2) + c_2.$$

Here we require that $|c_1|, |c_2| = o(1)$, i.e., the amplitudes of the perturbations are small as the common perturbation analysis (Bonnans and Shapiro 2013). We denote the total variation bound, i.e., the right-hand side of Eqn. (4) under $\{\tilde{f}_i\}_{i=1}^2$ and the corresponding $\{\tilde{G}_i^*\}_{i=1}^2$ as $\mathtt{TVB}(c_1, c_2)$.

Proposition 2 (Annealing Property of the Relaxation Criterion)**.** *Maintaining the same expectation of the number of accepted tokens for $\{\tilde{f}_i\}_{i=1}^2$ and $\{f_i\}_{i=1}^2$ requires that*

$$c_1 = \frac{-\mathbb{E}_{X_1 \sim Q}\big[f_1(X_1, P, Q; \omega_i)\big] \cdot c_2}{1 + \mathbb{E}_{(X_1, X_2) \sim Q}\Big[f_2\big((X_1, X_2), P, Q; \omega_i\big)\Big] + c_2}. \quad (8)$$

If some regularties conditions hold and $f_2^{\mathrm{van}}((x_1, x_2), P, Q) \leq f_2((x_1, x_2), P, Q; \omega_i)$ pointwisely, we have the following for two perturbations (c_1, c_2) and $(\tilde{c}_1, \tilde{c}_2)$ satisfying Eqn. (8) *but with different signs, i.e., $c_1, \tilde{c}_2 > 0$, $c_2, \tilde{c}_1 < 0$,*

$$\mathtt{TVB}(c_1, c_2) < \mathtt{TVB}(\tilde{c}_1, \tilde{c}_2). \quad (9)$$

The mentioned regularity conditions and the formal statement of this proposition are provided in Appendix F. This proposition indicates that increasing relaxation at an earlier position while reducing it at a later one—as specified in (8)—leads to lower distributional bias than the reverse, i.e., (9). When $L > 2$, a similar conclusion can be drawn by comparing each pair of neighboring positions.

Therefore, we propose an *annealed relaxation of speculative decoding* (**COOL-SD**) scheme for the acceptance criterion (see Algorithm 1). Instead of using a fixed relaxation parameter across all drafted positions, our approach progressively tightens the acceptance condition as decoding proceeds, i.e., $\omega_0 > \omega_1 > \ldots > \omega_L$. Specifically, we define the ω_i in Eqn. (7) to be

$$\text{(COOL-SD)} \qquad \omega_i = \delta \cdot \exp(-\nu \cdot i - \mu), \quad (10)$$

for all $i \in [L]$ with the decaying coefficient $\nu > 0$ and normalizing parameter μ such that $\sum_{i=1}^L \exp(-\nu \cdot i - \mu) = L$. COOL-SD is expected to induce less shift than UNIFORMRSD under the same accepted length. In Section 5, we show that empirically COOL-SD achieves better image quality than UNIFORMRSD at the same inference speed.

Target Model	Method	CLIP (↑)	FID (↓)	IR (↑)	Acc. Len. (↑)	Latency/s (↓)	Speed-up/× (↑)
Lumina-mGPT	Target Model	0.3330	28.99	0.6855	1.00 (±0.00)	170.14 (±1.32)	1.00
	Eagle-1	0.3330	29.05	0.6883	2.76 (±0.07)	71.66 (±1.80)	2.37
	LANTERN++ (λ=2, k=10)	**0.3328**	30.31	0.6697	2.99 (±0.07)	68.64 (±1.86)	2.48
	COOL-SD (δ=1.1)	0.3325	**30.30**	**0.6699**	**3.11** (±0.07)	**63.24** (±1.83)	**2.69**
LlamaGen-XL	Target Model	0.3162	21.08	-0.0763	1.00 (±0.00)	10.11 (±0.82)	1.00
	Eagle-1	0.3157	20.97	-0.0859	2.42 (±0.15)	4.99 (±0.34)	2.03
	LANTERN++ (λ=2, k=10)	0.3157	21.17	-0.1155	2.67 (±0.18)	4.70 (±0.38)	2.15
	COOL-SD (δ=1.1)	**0.3167**	**21.02**	**-0.0997**	2.73 (±0.16)	4.46 (±0.34)	2.27
	COOL-SD (δ=2)	0.3154	21.20	-0.1353	**3.34** (±0.20)	**3.72** (±0.27)	**2.72**

Table 1: Quantitative results with Lumina-mGPT and LlamaGen-XL as the target models. The bold indicates the best results among the SD methods that consider relaxations. COOL-SD achieves a better trade-off between latency and generation quality. The parameters λ and k are relaxation hyperparameters used in LANTERN++; we adopt the same values as in the experiments reported by Park et al. (2025). Standard deviations are shown in parentheses.

4 Experiments

In this section, we evaluate our method through quantitative and qualitative results, comparing it with vanilla SD and LANTERN++. We also show visual results under different relaxation budgets δ to show the latency-quality trade-off.

4.1 Implementation Details

We conduct extensive experiments to assess the effectiveness of COOL-SD on LlamaGen-XL (Sun et al. 2024a) (775M) and Lumina-mGPT (Liu et al. 2024) (7B) for accelerating visual AR inference, using a single NVIDIA A100-SXM4-40GB GPU. See Appendix C for more details.

Draft Model Training and Inference For training the draft models, we follow a pipeline similar to Eagle (Li et al. 2024b) and LANTERN++ (Park et al. 2025). We evaluate all methods on 5k randomly selected captions from the MS-COCO 2017 validation set (Lin et al. 2014), following the inference setup of LANTERN (Jang et al. 2024) and LANTERN++ (Park et al. 2025).

Metrics We measure acceleration by average accepted length and latency, where the former is the ratio of generated tokens to speculative decoding rounds. Generation quality is evaluated using the Fréchet Inception Distance (FID) (Heusel et al. 2017), CLIP score (Hessel et al. 2021), and ImageReward (IR) (Xu et al. 2023).

Baselines We benchmark our method against Eagle-1 (Li et al. 2024b), which serves as a baseline for lossless speculative decoding, and LANTERN++ (Park et al. 2025), a state-of-the-art relaxation-based speculative decoding approach. We adopt the static tree structure in Eagle-1 (Li et al. 2024b), which has been demonstrated by Park et al. (2025) to outperform the dynamic tree structure (Li et al. 2024a).

4.2 Experimental Results

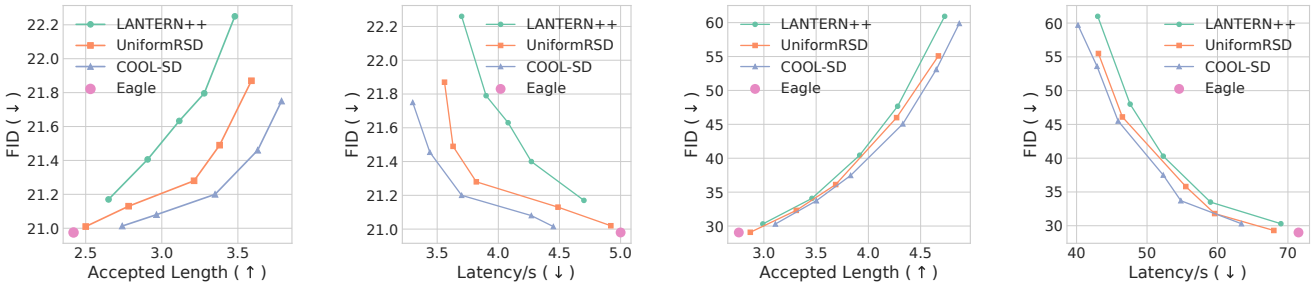
Quantitative Results Table 1 summarizes our main results. Compared to the lossless Eagle-1, COOL-SD achieves significantly faster inference with only minor quality degradation. The speed-quality trade-off in COOL-SD is tunable via the relaxation budget δ: smaller δ maintains near-baseline quality with faster inference, while larger δ offers greater speedups with modest quality loss.

Specifically, on LlamaGen, our method increases the average number of accepted steps from 2.42 to 2.73 with negligible loss in FID and CLIP Score. When the accepted length is further raised to 3.34, the image quality remains nearly unchanged. Similarly, on Lumina-mGPT, we observe consistent improvements: the accepted length rises from 2.75 to 3.15, leading to a reduction of approximately 10 seconds per image on the latency, while the FID and CLIP Score degradation remains within an acceptable range.

Compared to the relaxation-based SD method LANTERN++, COOL-SD consistently produces higher-quality images under the same latency constraints or achieves greater acceleration at similar quality, showing superior efficiency. For a fair comparison with the hyperparameters $\lambda = 2, k = 10$ used in the LANTERN++ paper, we set the hyperparameter δ of COOL-SD to 1.1 or 2.0, and ν to 0.7. As shown in Figure 3 of the next section, even when varying δ for COOL-SD and λ, k for LANTERN++, COOL-SD still outperforms LANTERN++ across settings. Additional comparisons with Speculative Jacobi Decoding (SJD) (Teng et al. 2024) are presented in Appendix C.

Qualitative Results In Figure 2, we present visualizations of our method under different choices of relaxation budget δ from 1.1 to 3.0. Increasing δ effectively improves the inference speed, demonstrating that the trade-off between speed and generation quality can be flexibly controlled through this parameter. As shown in the figure, the visual quality remains comparable to vanilla SD even as the inference speed increases from $\times 2.37$ to $\times 3.10$. Although further increasing the inference speed to $\times 3.70$ and beyond may result in some degradation in image quality, this trade-off remains easily manageable and can be tuned to meet practical needs.

We also emphasize that our method achieves superior speed-quality trade-off compared to LANTERN++ (Park et al. 2025). The corresponding quantitative results are presented in Section 5, while qualitative comparisons with LANTERN++ are provided in Appendix C.



(a) FID v.s. Accepted Length on LlamaGen-XL.

(b) FID v.s. Latency on LlamaGen-XL.

(c) FID v.s. Accepted Length on Lumina-mGPT.

(d) FID v.s. Latency on Lumina-mGPT.

Figure 3: The trade-off curves between imaging quality (evaluated by FID) and accepted length/lantency. We compared COOL-SD with UNIFORMRSD and LANTERN++ with $k = 10$ on two target models: LlamaGen-XL and Lumina-mGPT. We tested on 5000 random sampled captions from MSCOCO validation set.

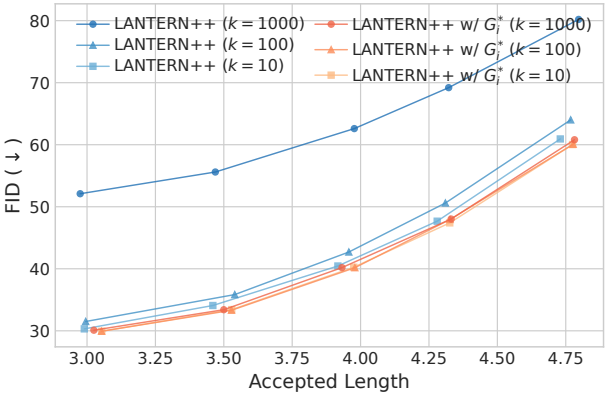


Figure 4: Ablation study on LANTERN++ with our resampling distribution G_i^*, under different settings of k.

5 Ablation Study

This section focuses on ablation studies across a wide range of parameter settings for COOL-SD, comparing it with LANTERN++. We also analyze the impact of our resampling distribution when applied to LANTERN++.

5.1 Trade-off Between Speed and Quality

In this paper, we propose a principled approach COOL-SD to control the trade-off between inference speed and image quality by adjusting δ in the range from 1.1 to 4.0, and $\nu = 0.7$ in COOL-SD. Figure 3 illustrates the relationship between FID and accepted length/latency for the two target models, LlamaGen-XL (Sun et al. 2024a) and Lumina-mGPT (Liu et al. 2024). Evaluating different methods requires examining the trade-off between generation quality and acceleration. As shown in Figures 3(a) and (b), for the same FID level (i.e., along a horizontal slice of the plot), COOL-SD achieves a greater accepted length and lower latency. Conversely, at a fixed acceptance length or latency, COOL-SD attains a lower FID compared to LANTERN++. Moreover, due to its simple implementation—free from the nearest-token probability retrieval and summation required by LANTERN++—COOL-SD achieves additional latency improvements, a decisive factor for real-world applications.

We further compare COOL-SD with UNIFORMRSD, a straightforward setting in which all drafting steps share the same relaxation parameter, as shown in Figure 3. The results demonstrate that the annealing strategy significantly enhances relaxation performance, yielding better trade-off between generation quality and inference speed.

5.2 Principled Resampling Strategy

LANTERN (Jang et al. 2024) and LANTERN++ (Park et al. 2025) proposed to perform relaxation on SD based on image token similarity. However, a mathematically rigorous proof is missing for their corresponding resampling distribution. To illustrate the importance of the resampling distribution and the effectiveness of our method, we conduct an experiment on LANTERN++ and LANTERN++ with our resampling distribution G^* in Eqn. (5). Our experiments are conducted under different settings of k—the hyperparameter in LANTERN++ that controls the number of nearest tokens in the latent space considered as similar. The result is shown in Figure 4. As demonstrated, with our resampling distribution, LANTERN++'s performance is improved regardless of the settings of hyperparameter k. Also, since LANTERN++ adopts an inferior resampling strategy, as k is set larger, the bias introduced by their algorithm grows. This issue is empirically mitigated by our resampling distribution, as evidenced by the orange lines in Figure 4.

6 Conclusion

In this paper, we introduce COOL-SD, a principled relaxed SD framework that consistently improves the speed–quality trade-off in auto-regressive image generation through two key insights: (1) establishing a tight bound on distributional bias and deriving token-wise optimal resampling, and (2) developing a provably effective decaying acceptance schedule. Extensive experiments demonstrate that COOL-SD outperforms existing speculative decoding variants in both efficiency and fidelity.

Acknowledgments

Xingyao Li acknowledges the support of Sea AI Lab for providing computational resources. This work is also supported by the Singapore Ministry of Education Academic Research Fund (AcRF) Tier 1 Grant (No. A-8000981-00-00).

References

Bachmann, G.; Anagnostidis, S.; Pumarola, A.; Georgopoulos, M.; Sanakoyeu, A.; Du, Y.; Schönfeld, E.; Thabet, A.; and Kohler, J. 2025. Judge decoding: Faster speculative sampling requires going beyond model alignment. *arXiv preprint arXiv:2501.19309*.

Bonnans, J. F.; and Shapiro, A. 2013. *Perturbation analysis of optimization problems*. Springer Science & Business Media.

Cai, T.; Li, Y.; Geng, Z.; Peng, H.; Lee, J. D.; Chen, D.; and Dao, T. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

Chen, C.; Borgeaud, S.; Irving, G.; Lespiau, J.-B.; Sifre, L.; and Jumper, J. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Du, C.; Jiang, J.; Yuanchen, X.; Wu, J.; Yu, S.; Li, Y.; Li, S.; Xu, K.; Nie, L.; Tu, Z.; et al. 2024. Glide with a cape: A low-hassle method to accelerate speculative decoding. *arXiv preprint arXiv:2402.02082*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

He, Y.; Chen, F.; He, Y.; He, S.; Zhou, H.; Zhang, K.; and Zhuang, B. 2025. ZipAR: Parallel Autoregressive Image Generation through Spatial Locality. In *Forty-second International Conference on Machine Learning*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, Z.; and Huang, H. 2024. Accelerated Speculative Sampling Based on Tree Monte Carlo. In *Forty-first International Conference on Machine Learning*.

Jang, D.; Park, S.; Yang, J. Y.; Jung, Y.; Yun, J.; Kundu, S.; Kim, S.-Y.; and Yang, E. 2024. Lantern: Accelerating visual autoregressive models with relaxed speculative decoding. *arXiv preprint arXiv:2410.03355*.

Leviathan, Y.; Kalman, M.; and Matias, Y. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, 19274–19286. PMLR.

Li, Y.; Wei, F.; Zhang, C.; and Zhang, H. 2024a. EAGLE-2: Faster Inference of Language Models with Dynamic Draft Trees. In *Empirical Methods in Natural Language Processing*.

Li, Y.; Wei, F.; Zhang, C.; and Zhang, H. 2024b. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, D.; Zhao, S.; Zhuo, L.; Lin, W.; Xin, Y.; Li, X.; Qin, Q.; Qiao, Y.; Li, H.; and Gao, P. 2024. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*.

Miao, X.; Oliaro, G.; Zhang, Z.; Cheng, X.; Wang, Z.; Wong, R. Y. Y.; Chen, Z.; Arfeen, D.; Abhyankar, R.; and Jia, Z. 2023. SpecInfer: Accelerating Generative LLM Serving with Speculative Inference and Token Tree Verification. *ArXiv*, abs/2305.09781.

Pang, Y.; Jin, P.; Yang, S.; Lin, B.; Zhu, B.; Tang, Z.; Chen, L.; Tay, F. E.; Lim, S.-N.; Yang, H.; et al. 2024. Next patch prediction for autoregressive visual generation. *arXiv preprint arXiv:2412.15321*.

Park, S.; Jang, D.; Kim, S.; Kundu, S.; and Yang, E. 2025. LANTERN++: Enhancing Relaxed Speculative Decoding with Static Tree Drafting for Visual Auto-regressive Models. *arXiv preprint arXiv:2502.06352*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schuhmann, C.; Köpf, A.; Coombes, T.; Vencu, R.; Trom, B.; and Beaumont, R. 2022. LAION COCO: 600M SYNTHETIC CAPTIONS FROM LAION2B-EN. https://huggingface.co/datasets/laion/laion-coco. Accessed: 2025-02-20.

Sun, P.; Jiang, Y.; Chen, S.; Zhang, S.; Peng, B.; Luo, P.; and Yuan, Z. 2024a. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Sun, Z.; Mendlovic, U.; Leviathan, Y.; Aharoni, A.; Beirami, A.; Ro, J. H.; and Suresh, A. T. 2024b. Block Verification Accelerates Speculative Decoding. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*.

Sun, Z.; Suresh, A. T.; Ro, J. H.; Beirami, A.; Jain, H.; and Yu, F. 2024c. Spectr: Fast speculative decoding via optimal transport. *Advances in Neural Information Processing Systems*, 36.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Team, C. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Teng, Y.; Shi, H.; Liu, X.; Ning, X.; Dai, G.; Wang, Y.; Li, Z.; and Liu, X. 2024. Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding. *arXiv preprint arXiv:2410.01699*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; et al. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Wang, Y.; Ren, S.; Lin, Z.; Han, Y.; Guo, H.; Yang, Z.; Zou, D.; Feng, J.; and Liu, X. 2025. Parallelized autoregressive visual generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12955–12965.

Xie, J.; Mao, W.; Bai, Z.; Zhang, D. J.; Wang, W.; Lin, K. Q.; Gu, Y.; Chen, Z.; Yang, Z.; and Shou, M. Z. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Xiong, J.; Liu, G.; Huang, L.; Wu, C.; Wu, T.; Mu, Y.; Yao, Y.; Shen, H.; Wan, Z.; Huang, J.; et al. 2024. Autoregressive models in vision: A survey. *arXiv preprint arXiv:2411.05902*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935.

Yin, M.; Chen, M.; Huang, K.; and Wang, M. 2024. A theoretical perspective for speculative decoding algorithm. *arXiv preprint arXiv:2411.00841*.

Zhong, M.; Teku, N.; and Tandon, R. 2025. Speeding up Speculative Decoding via Sequential Approximate Verification. *arXiv preprint arXiv:2502.04557*.