Continuous Speculative Decoding for Autoregressive Image Generation

Zili Wang[1,2], Zheng Zhang[3], Kun Ding[1,2], Qi Yang[1,2], Fei Li[4], and Shiming Xiang[1,2⋆]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[2] MAIS, Institute of Automation, Chinese Academy of Sciences, China
smxiang@nlpr.ia.ac.cn
[3] JD.COM Inc, China
[4] China Tower Corporation Limited, China

Fig. 1: Continuous speculative decoding accelerates the inference speed while maintaining the generation quality. For each panel, left: image generated by MAR; right: image generated by MAR with continuous speculative decoding with speed-up ratio.

Abstract. Continuous visual autoregressive (AR) models have demonstrated promising performance in image generation, but their inherently sequential nature results in slow inference speed. Speculative decoding, a successful acceleration technique for large language models (LLMs), has effectively accelerated discrete visual AR models. However, the absence of an analogous theory for continuous distributions precludes its use in accelerating continuous AR models. To fill this gap, this work presents continuous speculative decoding, and addresses challenges from: 1) low acceptance rate, caused by inconsistent output distribution modeled by target and draft models, and 2) modified distribution without analytic expression, caused by a complex integral. For challenge 1), we address low acceptance rates through an approximated criterion, a novel denoising trajectory alignment strategy based on reparameterization proximity, and token pre-filling. For challenge 2), we introduce acceptance-rejection

⋆ Corresponding author.

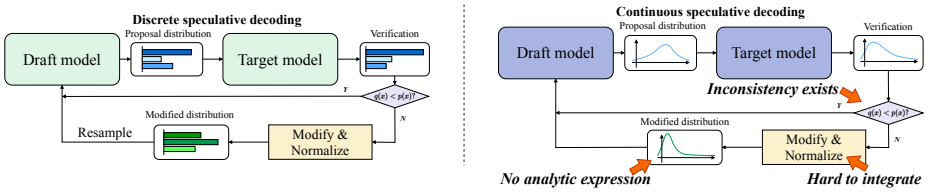

Fig. 2: Comparison between discrete and continuous speculative decoding. Discrete situation offers the convenience of directly computing probabilities and simply sampling from modified distributions. In contrast, continuous situation faces challenges in the inconsistency of output distributions, leading to low acceptance criterion as well as low acceptance rate, and the modified distributions without analytic expression, caused by complex integral.

sampling algorithm with an appropriate upper bound, thereby avoiding explicitly calculating the integral. Furthermore, our denoising trajectory alignment is also reused in acceptance-rejection sampling, effectively avoiding repetitive diffusion model inference. Extensive experiments on various models at 256×256 and 512×512 resolutions demonstrate that our approach achieves over 2× wall-time speedup while preserving the image generation quality.

Keywords: Visual autoregressive models · Image generation · Speculative decoding · Diffusion models

1 Introduction

Autoregressive (AR) models have demonstrated significant potential and achieved competitive performance in image generation tasks [7,27,37,39,43]. Existing approaches operate either in discrete token spaces using vector quantization [24,43] or in continuous latent spaces with a diffusion model [17, 33, 38, 41]. Discrete AR models suffer from training instability and reconstruction degradation introduced by vector quantization [24, 43]. In contrast, continuous visual AR models predict continuous latent variables via denoising-based transitions conditioned on previous tokens [11,26], overcome the issues of the discrete counterparts and offer a promising solution for autoregressive image generation. However, similar to LLMs, continuous visual AR models remain slow at inference due to their inherently sequential decoding process.

To accelerate autoregressive inference, speculative decoding [3,15] has proven effective for LLMs. This algorithm employs a draft-and-verification mechanism, where a smaller draft model generates several draft tokens, and which are then verified in parallel by a larger and more powerful target model. The acceptance of each draft token depends on the probability ratio (likelihood ratio) between the target and draft model distributions. Tokens with higher ratios are more likely to be accepted. Recent work has extended speculative decoding to discrete

visual AR models [12, 36]. However, these methods assume discrete categorical output distributions and do not directly extend to continuous domains.

This paper introduces **Continuous Speculative Decoding**, a novel framework to accelerate inference for continuous visual AR models. Applying speculative decoding to continuous distributions presents two core challenges, as illustrated in Fig. 2: a) **Low acceptance rate.** the draft and target models often learn divergent data distributions. As a result, samples generated by the draft model may lie outside high-probability regions under the target model, leading to low probability ratios and low acceptance rates. b) **Non-analytic modified distribution.** When a proposed token is rejected, a new token will be drawn from the modified distribution. In continuous distributions, this distribution lacks an analytic expression due to a complex normalizing integral, making direct sampling intractable.

To overcome these challenges, we propose a set of synergistic solutions. **First**, we address the low acceptance rate via a three-pronged approach: (i) an *approximated acceptance criterion* computed using the probabilities of the samples drawn from the draft model and the target model, respectively, enabling efficient sampling; (ii) a *denoising trajectory alignment* method, grounded in a theory of *reparameterization proximity*, to minimize the distributional divergence while reducing the bias introduced by the approximated acceptance criterion; and (iii) a token pre-filling strategy to stabilize acceptance rates in early generation steps. **Second**, to sample from the non-analytic modified distribution, we employ an acceptance-rejection sampling scheme [2]. We derive an appropriate upper bound to avoid computing the complex integral and introduce an easy-to-compute rejection threshold by reusing the property of denoising trajectory alignment, thereby avoiding extra model inference.

Our continuous speculative decoding can be integrated seamlessly into many existing models, as shown in Fig. 1. We validate the effectiveness of our algorithm on various continuous visual AR models [17, 33, 40] at two resolutions (256 & 512) through qualitative and quantitative evaluations. Specifically, we measure wall-time improvements and report image generation quality using different evaluation metrics, including Fréchet Inception Distance (FID) [10] and Inception Score (IS) [30]. Extensive experiments show that our algorithm achieves over 2× inference speedup while maintaining generation quality.

Our contributions can be summarized as follows:

- We are the first to propose continuous speculative decoding, bridging the gap of speculative decoding to continuous distributions and enabling substantial acceleration of continuous visual AR models.
- We address the problem of low acceptance rate in continuous speculative decoding via a novel approximated criterion, denoising trajectory alignment, and token pre-filling.
- We enable tractable sampling from the modified distribution via a tailored acceptance-rejection sampling scheme with a proper upper bound.
- We validate our proposed method into three existing continuous visual AR models without extra training or architectural changes. Extensive experi-

ments show that it achieves over 2× inference speedup while fully maintaining generation quality.

2 Related Work

2.1 Autoregressive Image Generation

Autoregressive (AR) models are widely used in image generation. Early works perform generation at pixel level using CNN [5,27], RNN [39] and Transformer [4, 29]. VAR [37] modifies the autoregressive paradigm into next-scale prediction to gradually increase the scale of predictions. Similar to the autoregressive language model, AR image generation through discrete token prediction is scalable to text-conditioned image generation [22, 34, 44]. However, training discrete image tokenizer is difficult, and its ability to convey detailed visuals is still questionable [24, 43]. GIVT [38] represents continuous tokens via Gaussian mixture models. MAR [17] and DisCo-Diff [41] generate tokens via diffusion process [11] conditioned by the autoregressive model. HART [35] employs discrete and continuous tokenizer to generate images, with classification for discrete tokens and denoising for the residual between primitive visual tokens and discrete tokens. xAR [33] extends the conception of token and reformulates discrete token classification as continuous entity regression. However, autoregressive models suffer from heavy inference overhead. The inference speed is slowed down by step-by-step generation.

2.2 Speculative Decoding

Speculative decoding [3, 15] achieves lossless acceleration by verifying the draft model with the target model. Following this, previous works mainly focus on reducing draft model overhead and strengthening the consistency between the draft and target models. SpecInfer [25] employs multiple small draft models and aggregates their predictions into a tree structure to be verified through tree-based parallel decoding. Eagle [19, 20] improves the draft accuracy through the prediction at the feature level instead of the token level to tackle the feature uncertainty problem. Jacobi iteration is also employed to reduce inference overhead in the decoding process [14,31,45,46]. Online Speculative Decoding [23] and DistillSpec [47] align the output from the draft model with the target model with more training. Speculative decoding can achieve lossless acceleration theoretically, but the generation quality may be affected under a larger speed-up ratio. BiLD [13] proposes a more relaxed acceptance condition. Besides, finetuning the target model can also improve generation quality [1, 14, 42].

Current literature explores applying speculative decoding to accelerate autoregressive image generation. Methods like SJD [36] combine speculative verification with Jacobi iterations to boost speed without sacrificing variety. Furthermore, LANTERN [12] and LANTERN++ [28] leverage relaxed candidate selection to manage distribution ambiguity and preserve image quality. Despite

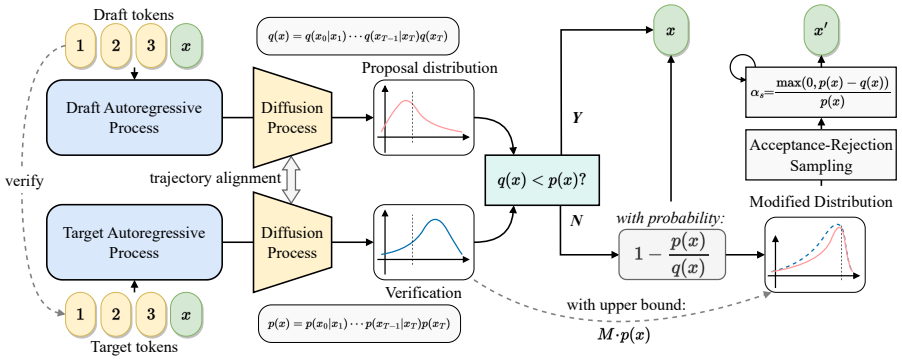

Fig. 3: The overview of continuous speculative decoding. The diffusion model component of continuous AR models is leveraged. Tokens $1 \sim 3$ are prefix tokens, and token x is to be verified. In this method, x is accepted if $q(x) < p(x)$. Otherwise, x is rejected with probability $1 - p(x)/q(x)$, followed by sampling x' from the modified distribution via acceptance-rejection sampling.

reducing total inference steps, these techniques are fundamentally limited to discrete token spaces. This paper addresses this limitation by proposing a continuous speculative decoding framework specifically designed for continuous autoregressive models.

3 Methodology

This section introduces our continuous speculative decoding framework for visual autoregressive models. We first extend discrete speculative decoding to continuous diffusion trajectories, then analyze why direct application leads to low acceptance rates. To mitigate this, we propose denoising trajectory alignment and target-token pre-filling to reduce draft-target inconsistency. Finally, we derive an efficient acceptance–rejection sampling strategy for resampling from the modified distribution.

3.1 From Discrete to Continuous Speculative Decoding

We first introduce the discrete form of speculative decoding. It utilizes a draft model M_q with output distribution $q(x)$ and a target model M_p with $p(x)$. M_q generates a sequence of draft tokens $x_{i:j} = \{x_i, \ldots, x_j\}$ where $x \sim q(x)$, which are then verified by M_p in parallel. x is accepted if the acceptance criterion $p(x)/q(x) > 1$; otherwise, it is rejected with probability $1 - p(x)/q(x)$ and resampled from a modified distribution $p'(x) = norm(max(0, p(x) - q(x))) = \frac{max(0,p(x)-q(x))}{\sum_{x'} max(0,p(x')-q(x'))}$.

Then, we discuss continuous speculative decoding (see Fig. 3). In continuous visual AR models [17], the output distribution is typically modeled via a diffusion

model [11, 26], namely:

$$p(x_{N,0:T}, |x_{1:N-1}) = p(x_{N,T}) \prod_{t=1}^{T} p(x_{N,t-1}|x_{N,t}, x_{1:N-1}), \tag{1}$$

where N represents N-th autoregressive step. $t \in [0, T]$ is the diffusion timestep. $x_{N,t}$ represents the state of x at autoregressive step N and diffusion timestep t. $x_{1:N-1} = \{x_1, x_2, \ldots, x_{N-1}\}$ denotes the variables before step N. For simplicity, we omit N and $x_{1:N-1}$ and let $Y = x_{0:T}$ to obtain $p(Y) = p(x_T) \prod_{t=1}^{T} p(x_{t-1}|x_t)$.

Acceptance Criterion. The acceptance criterion is defined as the ratio of the probability under target distribution to the one under draft distribution, that is, $p(x_{0:T})/q(x_{0:T})$. In discrete form, the probability can be directly obtained. But in continuous form, the probability is usually obtained via diffusion process [11, 32]. Specifically, $x_{0:T}$ is sampled from draft model via reverse diffusion (denoising) process. So the ratio is given by:

$$\frac{p(Y)}{q(Y)} = \frac{p(x_T) \prod_{t=1}^{T} p(x_{t-1}|x_t)}{q(x_T) \prod_{t=1}^{T} q(x_{t-1}|x_t)}, \tag{2}$$

where x_T is sampled from Gaussian noise and $p(x_{t-1}|x_t)$ is approximated as a Gaussian distribution through a neural network with parameter θ [26], and $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ are mean and variance predicted by θ, that is:

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{3}$$

However, this approach is impractical because the acceptance rate is low. As shown in Table 1 and discussed in Appendix A.1, the draft model's trajectory $x_{0:T}$ inherently diverges from the target model's expected trajectory. In each denoising step, samples drawn from the draft model's distribution q are unlikely to fall near μ of the target distribution p, which results in a low single-step ratio p/q. As the multi-step denoising process proceeds, the overall $p(Y)/q(Y)$ becomes extremely small.

Therefore, our work adopts a more practical approximation: the ratio of the joint probabilities $p(Y_p)/q(Y_q)$, where both Y_p and Y_q share the same x_0. This ratio serves as a surrogate for the shared path ratio. Then, the ratio is calculated through:

$$\frac{p(Y_p)}{q(Y_q)} \coloneqq \frac{p(x_T^p) \prod_{t=1}^{T} p(x_{t-1}^p|x_t^p)}{q(x_T^q) \prod_{t=1}^{T} q(x_{t-1}^q|x_t^q)}, \quad x_0^p = x_0^q = x_0. \tag{4}$$

Modified Distribution. When a draft token is rejected, a new token is resampled from the modified distribution $p'(x)$. Following the discrete formulation, $p'(x)$ is derived from the normalization of $max(0, p(x) - q(x))$. Therefore, replacing the summation in the normalization denominator of the discrete form with integral yields the continuous form, namely:

$$p'(Y) = \frac{max(0, p(Y) - q(Y))}{Z}, \text{ where } Z = \int_{Y'} max(0, p(Y') - q(Y'))dY'. \tag{5}$$

Table 1: Probability ratio of different calculation approaches. Y denotes the shared denoising trajectory of the sample. Y_p and Y_q represent the samples generated by the target model and the draft model through the denoising process, respectively.

Probability ratio	Value	Acceptance rate
$p(Y)/q(Y)$	5.33×10^{-23}	0.0%
$p(Y_p)/q(Y_q)$, w/o align	0.067	14%
$p(Y_p)/q(Y_q)$, w/ align	1.86	32%

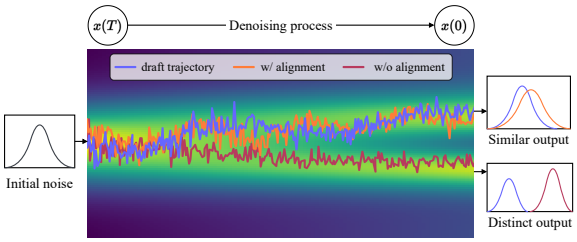

Fig. 4: Illustration of denoising trajectory alignment. The denoising process maps the noise distribution to data distribution through gradual denoising. These denoising steps form a trajectory. The aligned trajectory (orange curve) leads to a similar output distribution, while the unaligned one (red curve) produces a far-away one, obtaining low $p(Y_p)/q(Y_q)$.

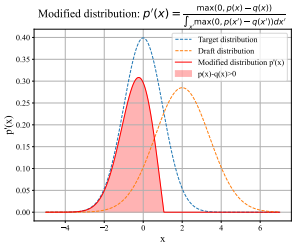

Fig. 5: Illustration of distributions. Dashed lines: draft and target distributions. Red area: modified distribution (unnormalized, omitting Z for simplicity), whose integral lack an analytic expression.

However, in practical operation, an extremely low acceptance rate emerges, as shown in Table 8, thereby leading to poor acceleration performance. To address this, we propose the following methods.

3.2 Mitigating distribution inconsistency

Under the acceptance criterion given by Equation 4, the acceptance rate α is extremely small (nearly 0%). We attribute this to two sources of inconsistency in continuous visual AR models: **1) inconsistency in diffusion process**, and **2) inconsistency in autoregressive process**.

Inconsistency in Diffusion Process. Significant inconsistency exists in the denoising process. As illustrated in Fig. 4, using the approximated acceptance criterion, draft and target denoising trajectories diverge to different outputs. The output distance between the two distributions is large. Besides, the approximated acceptance criterion is also biased relative to the original probability ratio, and therefore it does not preserve the unbiasedness guarantee of the original speculative decoding.

We propose denoising trajectory alignment to enhance the consistency of the output distributions and reduce the degree of bias. Note that in Equation 3, x_{t-1}^p

is obtained via reparameterization given by $x_{t-1}^p = \sqrt{\Sigma_\theta^p(x_t^p, t)} \cdot \varepsilon_t^p + \mu_\theta^p(x_t^p, t)$, where $\varepsilon_t^p \sim \mathcal{N}(0, \mathrm{I})$ (same for x_{t-1}^q). Denoising trajectory alignment can reduce the expected distance between x_{t-1}^p and x_{t-1}^q, promised by the following theorem.

Theorem 1 (Reparameterization Proximity). *Setting $\varepsilon_t^p = \varepsilon_t^q$ in reparameterization reduces the expected distance $\mathbb{E}\left[\left\|x_{t-1}^q - x_{t-1}^p\right\|^2\right]$ between x_{t-1}^p and x_{t-1}^q by $2 \cdot tr\left[\sqrt{\Sigma_t^q \Sigma_t^p}\right]$.*

Detailed proofs can be found in Appendix A. Note that:

$$\begin{aligned} p(x_{t-1}|x_t) &= \frac{\exp\{\frac{1}{2}[x_{t-1} - \mu_\theta(x_t,t)]^T \Sigma_\theta^{-1}(x_t,t)[x_{t-1} - \mu_\theta(x_t,t)]\}}{(\sqrt{2\pi})^n \sqrt{|\Sigma_\theta(x_t,t)|}} \\ &= \frac{1}{(\sqrt{2\pi})^n \sqrt{|\Sigma_\theta(x_t,t)|}} \exp\{\frac{1}{2}{\varepsilon_t}^T \varepsilon_t\}, \quad x_t \in \{x_t^p, x_t^q\}, \varepsilon_t \in \{\varepsilon_t^p, \varepsilon_t^q\}. \end{aligned} \tag{6}$$

The ratio $p(x_{t-1}^p|x_t^p)/q(x_{t-1}^q|x_t^q)$ given $\varepsilon_t = \varepsilon_t^p = \varepsilon_t^q$ is:

$$\frac{p(x_{t-1}^p|x_t^p)}{q(x_{t-1}^q|x_t^q)} = \frac{\frac{1}{(\sqrt{2\pi})^n \sqrt{|\Sigma_t^p|}} \exp\{\frac{1}{2}\varepsilon_t^T \varepsilon_t\}}{\frac{1}{(\sqrt{2\pi})^n \sqrt{|\Sigma_t^q|}} \exp\{\frac{1}{2}\varepsilon_t^T \varepsilon_t\}} = \frac{\sqrt{|\Sigma_t^q|}}{\sqrt{|\Sigma_t^p|}}. \tag{7}$$

For simplicity, we define:

$$\Sigma = \prod_{t=2}^{T} \sqrt{|\Sigma_t^q|} / \prod_{t=2}^{T} \sqrt{|\Sigma_t^p|}. \tag{8}$$

Substituting Σ into Equation 4 makes (assuming $p(x_T^p) = q(x_T^q)$):

$$\frac{p(Y_p)}{q(Y_q)} = \frac{p(x_T^p) p(x_0|x_1^p) \prod_{t=2}^{T} p(x_{t-1}^p|x_t^p)}{q(x_T^q) q(x_0|x_1^q) \prod_{t=2}^{T} q(x_{t-1}^q|x_t^q)} = \frac{p_\theta(x_0|x_1^p)}{q_\theta(x_0|x_1^q)} \cdot \Sigma. \tag{9}$$

Therefore, at each step, employing the same ε_t reduces the expected distance by $2 \cdot \mathrm{tr}\left[\Sigma_t^q \Sigma_t^p\right]$, which enables the two models to generate closer samples and finally increases the acceptance rate, as shown in Table 1.

Inconsistency in Autoregressive Process. Inconsistency also arises in autoregressive steps. Fig. 9 shows that acceptance rate α is very low (5%) at the initial AR steps, and it increases progressively as AR steps grow. This stems from the different draft and target prefix embeddings [17]. Owing to this, the AR models naturally yield divergent predictions, which in turn leads to low α. As the AR steps increase, the inputs of the two models gradually converge, thereby improving consistency between their outputs and finally raising the α.

To address this, we propose pre-filling a portion (e.g., 5%) of tokens from the target model to ensure a consistent prefix. This does not increase inference latency, as speculative decoding at a low acceptance rate is functionally equivalent to the target model step-by-step decoding [15]. Furthermore, pre-filling improves the overall acceptance rate.

Finally, $p(Y_p)/q(Y_q)$ can be computed with the help of denoising trajectory alignment and token pre-filling to obtain a considerable acceptance rate.

3.3 Resample from the modified distribution

The simplified illustration of Equation 5 is shown in Fig. 5. Unlike discrete form, where $\sum max(0, p(x) - q(x))$ can be directly computed, the analytic expression of Z can't be computed because it involves the integral of the product of a series of Gaussian distributions given by Equation 4 and 3. Therefore, $p'(Y)$ cannot be directly sampled.

To tackle this problem, a viable approach is acceptance–rejection sampling [2], which first samples from a proposal distribution, and then calculates a predefined rejection threshold α_s and samples $\epsilon \sim U(0, 1)$. If $\epsilon < \alpha_s$, the sample is accepted, otherwise it is rejected and sampling from the proposal distribution will be repeated until the sample is accepted.

To realize this sampling method, we first define α_s as:

$$\alpha_s = \frac{p'(Y)}{M \cdot p(Y)}, \tag{10}$$

where $p(Y)$ is the target distribution, and M is the upper bound factor that holds $M \cdot p(Y) \geq p'(Y)$ for any Y. Given $max(0, p(Y) - q(Y)) \leq p(Y)$, M is set to $1/Z$ considering that:

$$p'(Y) = \frac{max(0, p(Y) - q(Y))}{Z} \leq \frac{p(Y)}{Z} \mapsto M \cdot p(Y). \tag{11}$$

We substitute $M = 1/Z$ into Equation 10 to eliminate Z:

$$\alpha_s = \frac{max(0, p(Y) - q(Y))/Z}{p(Y)/Z} = \frac{max(0, p(Y) - q(Y))}{p(Y)}. \tag{12}$$

Equation 12 gives the analytic expression of α_s without calculating Z. However, in its naive implementation, repetitive diffusion model inference is needed to sample $p(Y)$. It brings heavy extra overhead. To tackle this problem, denoising trajectory alignment is introduced to simplify Equation 12. Accordingly, we have derived the following corollary.

Corollary 1 (Easy-to-Compute Rejection Threshold) *With denoising trajectory alignment introduced from Equation 9, the rejection threshold α_s has the easy-to-compute form:*

$$\alpha_s = \frac{max(0, \Sigma \cdot p_\theta(x_0|x_1^p) - q_\theta(x_0|x_1^q))}{\Sigma \cdot p_\theta(x_0|x_1^p)}. \tag{13}$$

See Appendix A for detailed proofs. In this form, x_0 is sampled from $p_\theta(x_0|x_1^p)$, which is merely a Gaussian distribution defined by Equation 3, thus avoiding extra model inference. We can compute α_s by gathering Σ and sampling x_0 from the Gaussian distribution, thereby completing the acceptance-rejection sampling to obtain samples equivalent to those derived from sampling $p'(Y)$.

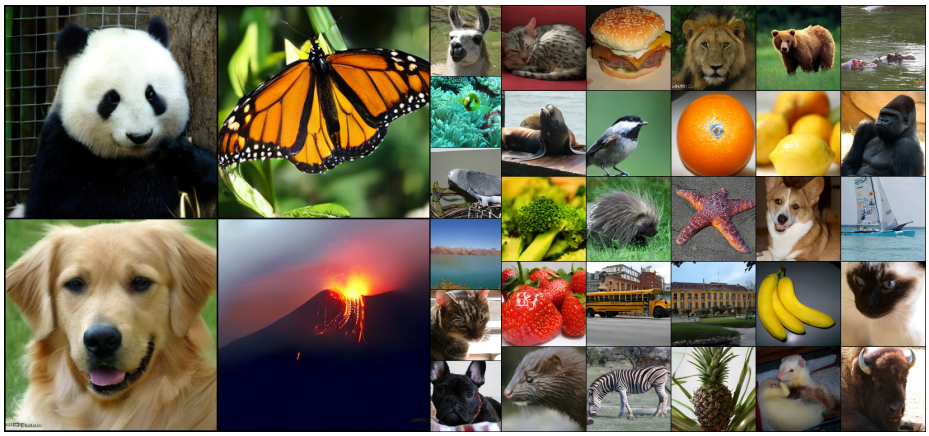

Fig. 6: Images generated by MAR with continuous speculative decoding.

Table 2: Results of speedup ratio and acceptance rate α on MAR and xAR under different draft lengths and batch sizes. The bs refers to batch size.

M_p	M_q	γ	α	Speedup ratio			
				bs=1	bs=8	bs=128	bs=256
MAR-H	MAR-B	32	0.19	**1.44**×	**1.61**×	**2.17**×	**2.33**×
MAR-H	MAR-B	16	0.26	1.37×	1.51×	2.07×	2.20×
MAR-H	MAR-B	8	0.27	1.26×	1.44×	1.88×	1.96×
MAR-H	MAR-B	4	0.30	1.11×	1.20×	1.56×	1.62×
xAR-H	xAR-B	32	0.22	**1.77**×	**2.10**×	**2.52**×	**2.72**×
xAR-H	xAR-B	16	0.26	1.58×	2.06×	2.31×	2.61×
xAR-H	xAR-B	8	0.29	1.61×	1.86×	2.07×	2.18×
xAR-H	xAR-B	4	0.36	1.30×	1.62×	1.92×	2.11×

4 Experiment

4.1 Implementation Details

We evaluate our method across several open-source continuous visual autoregressive models: MAR [17], xAR [33], and Harmon [40]. Experiments are conducted on ImageNet [6] for MAR and xAR at 256 × 256 resolution, while Harmon is tested at both 256 × 256 and 512 × 512 resolutions, leveraging its diverse pre-training data. We pair draft models—MAR-B (208M), xAR-B (172M), and Harmon-0.5B—with their corresponding target models: MAR-H (943M), xAR-H (1.1B), and Harmon-1.5B. Performance is reported using Fréchet Inception Distance (FID) [10] and Inception Score (IS) [30] for MAR and xAR, and FID, CLIPScore [9], and Geneval [8] for Harmon. Speedups are reported based on

Table 3: Results of speedup ratio and acceptance rate α on Harmon under different resolutions, draft lengths, and batch sizes (bs). Due to CUDA out-of-memory, this set of experiments opt for smaller batch sizes.

Resolution	M_p	M_q	γ	α	Speedup ratio			
					bs=1	bs=8	bs=16	bs=32
256	Harmon-H	Harmon-B	32	0.17	**1.47**×	**1.67**×	**1.88**×	**2.05**×
	Harmon-H	Harmon-B	16	0.21	1.41×	1.58×	1.78×	1.93×
	Harmon-H	Harmon-B	8	0.25	1.29×	1.44×	1.60×	1.72×
	Harmon-H	Harmon-B	4	0.33	1.11×	1.22×	1.33×	1.42×
512	Harmon-H	Harmon-B	32	0.15	**1.63**×	**1.94**×	**2.23**×	**2.54**×
	Harmon-H	Harmon-B	16	0.23	1.55×	1.83×	2.09×	2.35×
	Harmon-H	Harmon-B	8	0.25	1.41×	1.65×	1.85×	2.05×
	Harmon-H	Harmon-B	4	0.38	1.20×	1.37×	1.50×	1.63×

Table 4: Evaluation of FID and IS on unconditional and conditional generation, compared with original MAR-L and MAR-H models. Our method achieves acceleration while maintaining performance within a reasonable interval.

M_p	M_q	w/o CFG		w/ CFG	
		FID↓	IS↑	FID↓	IS↑
MAR-L		2.60	221.4	1.78	296.0
MAR-L	MAR-B	2.59±0.04	218.4±3.4	1.81±0.05	303.7±4.3
MAR-H		2.35	227.8	1.55	303.7
MAR-H	MAR-B	2.36±0.05	228.5±2.2	1.60±0.05	301.6±2.6
MAR-H	MAR-L	2.34±0.04	228.9±2.8	1.57±0.04	301.4±2.5

wall-clock time on a single NVIDIA A100 GPU. More details of experiment settings, ablation studies and quantitative results can be found in Appendix.

4.2 Main Results

Speedup results. As shown in Tables 2 and 3, the speedup ratio and acceptance rate exhibit a positive correlation with batch size, with draft lengths ranging from 8 to 32. This trend highlights the growing efficacy of speculative decoding in higher-throughput scenarios. Specifically, our algorithm achieves substantial speedups of 2.33× on MAR, 2.72× on xAR, and 2.54× on Harmon, demonstrating robust performance across diverse model architectures.

Table 5: Evaluation of FID and IS for xAR models. The results show that using different draft models maintains generation quality (FID and IS) within a stable range compared to the target models alone.

M_p	M_q	FID↓	IS↑
xAR-L		1.87	274.8
xAR-L	xAR-B	1.88±0.08	270.4±6.9
xAR-H		1.79	288.9
xAR-H	xAR-B	1.82±0.07	286.1±4.2
xAR-H	xAR-L	1.78±0.07	287.7±2.7

Table 6: Evaluation of FID and CLIPScore for Harmon with continuous speculative decoding on MSCOCO and MJHQ. The results show consistent performance between the standalone model and the speculative decoding setup.

M_p	M_q	FID		CLIPScore
		MSCOCO	MJHQ	MSCOCO
Harmon-H		8.39	5.15	34.8
Harmon-H	Harmon-B	8.38	5.13	34.7

Table 7: Harmon Evaluation of Geneval. The results demonstrate the performance across various fine-grained tasks, comparing the standalone Harmon-H model with the Harmon-H and Harmon-B combination.

M_p	M_q	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall
Harmon-H		0.99	0.86	0.66	0.85	0.74	0.48	0.76
Harmon-H	Harmon-B	0.99	0.83	0.66	0.86	0.74	0.44	0.75

Quantitative results. Tables 4 and 5 present the class-conditioned and unconditioned FID and IS metrics for our continuous speculative decoding applied to MAR and xAR. For Harmon, FID on MSCOCO [21] and MJHQ [16], alongside CLIPScore [9] and Geneval [8] are reported in Tables 6 and 7. We conduct multiple experiments and report the average performance and the standard deviation to ensure the reliability of our conclusions. These results demonstrate that our algorithm significantly preserves the quality of generated images, which will be further discussed in subsequent sections. Thus, our approach offers a robust solution for efficient and reliable model inference.

Qualitative results. Visual comparisons in Figures 6 and 7 demonstrate the superior image quality produced by our algorithm. While Figure 6 highlights the qualitative results of our method, Figure 7 provides a direct contrast with the baseline MAR-H (autoregressive steps=256) under varying draft lengths γ. These

Table 8: Ablation study on the acceptance rate α and the average distance of each draft and target token with and without denoising trajectory alignment under different draft length γ.

M_p	M_q	γ	α		$\mathbb{E}\left[\|x^q - x^p\|^2\right]$	
			w/o align	w/ align	w/o align	w/ align
MAR-H	MAR-B	32	0.07	**0.30**	2.56	**1.13**
MAR-H	MAR-B	16	0.07	**0.33**	2.36	**0.91**
MAR-H	MAR-B	8	0.13	**0.31**	2.22	**0.82**
MAR-H	MAR-B	4	0.14	**0.32**	2.17	**0.80**

Table 9: Ablation study on pre-filling ratio. Underline indicates the highest speedup. **Bold** means the highest α.

M_p	M_q	γ	α/Speed		
			0%	5%	15%
MAR-H	MAR-B	32	0.25/1.63×	0.30/1.63×	**0.33**/1.61×
MAR-H	MAR-B	16	0.32/1.53×	0.33/1.52×	**0.34**/1.51×
MAR-H	MAR-B	8	0.33/1.47×	0.31/1.47×	**0.34**/1.44×
MAR-H	MAR-B	4	0.31/1.21×	0.32/1.21×	**0.34**/1.20×

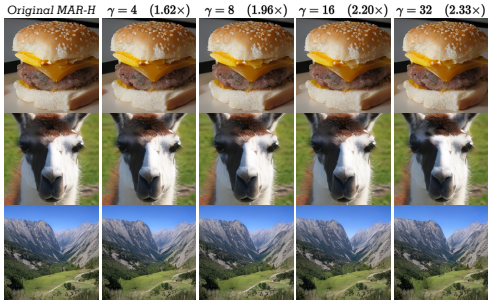

Fig. 7: Qualitative Comparison Results. We show the images using different draft length γ.

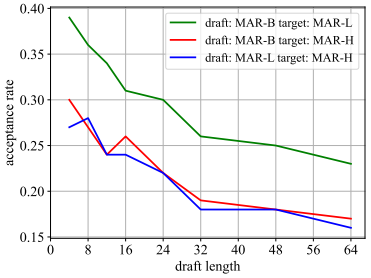

Fig. 8: Plots of acceptance rate α along with draft lengths γ on MAR.

visualizations confirm that our approach achieves significant acceleration while preserving image fidelity.

4.3 Ablation Study

The α vs. γ. The relationship between acceptance rate α and draft length γ on MAR model is depicted in Fig. 8. As the length of the draft increases, the acceptance rate tends to decline. This observation suggests that while longer drafts can substantially mitigate inference overhead, they are intrinsically constrained by the capabilities of the draft model itself. Consequently, an increase in the number of draft lengths is associated with greater deviations from the target model's distribution, ultimately leading to reduced acceptance rates.

Effectiveness of denoising trajectory alignment. Table 8 shows the acceptance rate and the average distance of each draft and target token with and without denoising trajectory alignment. The results show that, denoising trajectory alignment reduces the distance between the draft tokens and the target tokens. Before alignment, the distance between them reaches > 2, leading to quite small α. As the distance is reduced to ~ 1, the probability ratio increases, further increasing the α to $> 30\%$.

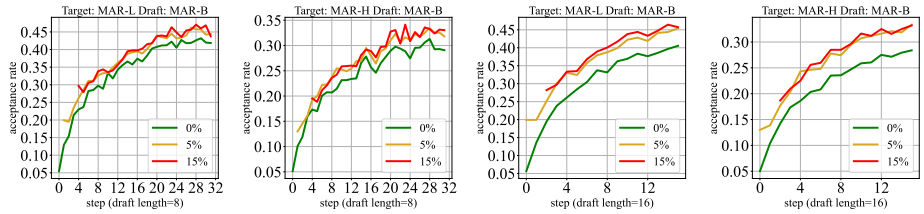

Fig. 9: Plots of per-step acceptance rate α under different pre-filling ratios, along with different draft length γ, averaged on 1000 samples from MAR.

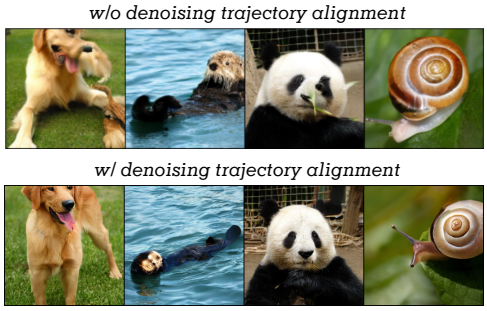

Fig. 10: The examples without (upper) and with (lower) denoising trajectory alignment. After alignment, the generated images exhibit a reduction in deformations and artifacts, thereby achieving higher quality.

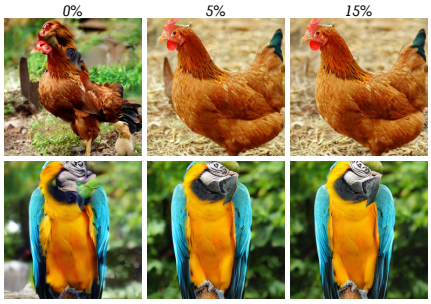

Fig. 11: Comparing image generation quality under different token pre-filling portions. The integration of token pre-filling mitigates anomalies and synthesis errors, enhancing visual fidelity.

Fig. 10 illustrates the image generated with and without alignment. Without denoising trajectory alignment, using trajectories sampled from the draft and target models independently leads to significant divergence, which not only reduces the token acceptance rate but also causes deformations and artifacts. This quality degradation is attributable to the fact that the approximate acceptance criterion is biased relative to the original probability ratio. With the application of denoising trajectory alignment, these negative effects are alleviated: divergence is significantly reduced, token acceptance rates are improved, and deformations and artifacts are effectively eliminated.

Influence of pre-filled tokens. The ablation study of pre-filling ratios at 0%, 5%, and 15% on MAR model is illustrated in Fig. 9. Pre-filling can compensate for the low acceptance rates observed during the initial stages of autoregressive sampling and enhance the overall acceptance rate, as shown in Table 9. Moreover, Fig. 11 shows the visualizations under different pre-filling ratios. Notably, the discrepancies between the draft and the target model result in certain artifacts and reduced image quality at 0% pre-filling. However, introducing a modest proportion of pre-filled tokens from the target model has effectively mitigated

these artifacts. As the pre-filling ratio increases, the advantages conferred by this approach exhibit diminishing returns.

5 Conclusion

We present continuous speculative decoding to accelerate continuous visual AR models. We propose denoising trajectory alignment based on proximity in the reparameterization theorem and token pre-filling to enhance the acceptance rate. Acceptance-rejection sampling is introduced with a proper upper bound to sample the modified distribution without analytic expression. The repetitive diffusion model inference is tackled by reusing denoising trajectory alignment. Extensive experiments show that our algorithm achieves over $2\times$ speedup while maintaining the output distribution. We expect our work will provide more thoughts and insights into the inference acceleration with continuous autoregressive models in vision and other domains.

References

1. Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J.D., Chen, D., Dao, T.: Medusa: Simple llm inference acceleration framework with multiple decoding heads. arXiv:2401.10774 (2024)
2. Casella, G., Robert, C.P., Wells, M.T.: Generalized accept-reject sampling schemes. Lecture notes-monograph series pp. 342–347 (2004)
3. Chen, C., Borgeaud, S., Irving, G., Lespiau, J.B., Sifre, L., Jumper, J.: Accelerating large language model decoding with speculative sampling. arXiv:2302.01318 (2023)
4. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning. pp. 1691–1703 (2020)
5. Chen, X., Mishra, N., Rohaninejad, M., Abbeel, P.: Pixelsnail: An improved autoregressive generative model. In: International Conference on Machine Learning. pp. 864–872 (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on Computer Vision and Pattern Recognition. pp. 248–255. Ieee (2009)
7. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
8. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
9. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems **30** (2017)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)

12. Jang, D., Park, S., Yang, J.Y., Jung, Y., Yun, J., Kundu, S., Kim, S.Y., Yang, E.: Lantern: Accelerating visual autoregressive models with relaxed speculative decoding. arXiv:2410.03355 (2024)
13. Kim, S., Mangalam, K., Moon, S., Malik, J., Mahoney, M.W., Gholami, A., Keutzer, K.: Speculative decoding with big little decoder. Advances in Neural Information Processing Systems **36** (2024)
14. Kou, S., Hu, L., He, Z., Deng, Z., Zhang, H.: Cllms: Consistency large language models. arXiv:2403.00835 (2024)
15. Leviathan, Y., Kalman, M., Matias, Y.: Fast inference from transformers via speculative decoding. In: International Conference on Machine Learning. pp. 19274–19286 (2023)
16. Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation (2024)
17. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2025)
18. Li, Y., Ge, Y., Ge, Y., Luo, P., Shan, Y.: Dicode: Diffusion-compressed deep tokens for autoregressive video generation with language models. arXiv:2412.04446 (2024)
19. Li, Y., Wei, F., Zhang, C., Zhang, H.: Eagle-2: Faster inference of language models with dynamic draft trees. arXiv:2406.16858 (2024)
20. Li, Y., Wei, F., Zhang, C., Zhang, H.: Eagle: Speculative sampling requires rethinking feature uncertainty. arXiv:2401.15077 (2024)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
22. Liu, D., Zhao, S., Zhuo, L., Lin, W., Qiao, Y., Li, H., Gao, P.: Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. arXiv:2408.02657 (2024)
23. Liu, X., Hu, L., Bailis, P., Cheung, A., Deng, Z., Stoica, I., Zhang, H.: Online speculative decoding. arXiv:2310.07177 (2023)
24. Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M.: Finite scalar quantization: Vq-vae made simple. arXiv:2309.15505 (2023)
25. Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., Zhang, Z., Wong, R.Y.Y., Zhu, A., Yang, L., Shi, X., et al.: Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification. arXiv:2305.09781 (2023)
26. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171 (2021)
27. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. Advances in Neural Information Processing Systems **29** (2016)
28. Park, S., Jang, D., Kim, S., Kundu, S., Yang, E.: Lantern++: Enhanced relaxed speculative decoding with static tree drafting for visual auto-regressive models. arXiv:2502.06352 (2025)
29. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International Conference on Machine Learning. pp. 4055–4064 (2018)
30. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in Neural Information Processing Systems **29** (2016)

31. Santilli, A., Severino, S., Postolache, E., Maiorca, V., Mancusi, M., Marin, R., Rodolà, E.: Accelerating transformer inference for translation via parallel decoding. arXiv:2305.10427 (2023)
32. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (2020)
33. Sucheng, R., Qihang, Y., Ju, H., Xiaohui, S., Alan, Y., Liang-Chieh, C.: Beyond next-token: Next-x prediction for autoregressive visual generation. arXiv preprint arXiv:2502.20388 (2025)
34. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv:2406.06525 (2024)
35. Tang, H., Wu, Y., Yang, S., Xie, E., Chen, J., Chen, J., Zhang, Z., Cai, H., Lu, Y., Han, S.: Hart: Efficient visual generation with hybrid autoregressive transformer. arXiv:2410.10812 (2024)
36. Teng, Y., Shi, H., Liu, X., Ning, X., Dai, G., Wang, Y., Li, Z., Liu, X.: Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding. arXiv:2410.01699 (2024)
37. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in Neural Information Processing Systems **37**, 84839–84865 (2025)
38. Tschannen, M., Eastwood, C., Mentzer, F.: Givt: Generative infinite-vocabulary transformers. In: European Conference on Computer Vision. pp. 292–309. Springer (2025)
39. Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International Conference on Machine Learning. pp. 1747–1756 (2016)
40. Wu, S., Zhang, W., Xu, L., Jin, S., Wu, Z., Tao, Q., Liu, W., Li, W., Loy, C.C.: Harmonizing visual representations for unified multimodal understanding and generation (2025), https://arxiv.org/abs/2503.21979
41. Xu, Y., Corso, G., Jaakkola, T., Vahdat, A., Kreis, K.: Disco-diff: Enhancing continuous diffusion models with discrete latents. In: International Conference on Machine Learning (2024)
42. Yi, H., Lin, F., Li, H., Ning, P., Yu, X., Xiao, R.: Generation meets verification: Accelerating large language model inference with smart parallel auto-correct decoding. arXiv:2402.11809 (2024)
43. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 10459–10469 (2023)
44. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al.: Language model beats diffusion–tokenizer is key to visual generation. arXiv:2310.05737 (2023)
45. Zhao, W., Huang, Y., Han, X., Xiao, C., Liu, Z., Sun, M.: Ouroboros: Speculative decoding with large model enhanced drafting. arXiv:2402.13720 (2024)
46. Zhao, Y., Xie, Z., Liang, C., Zhuang, C., Gu, J.: Lookahead: An inference acceleration framework for large language model with lossless generation accuracy. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 6344–6355 (2024)
47. Zhou, Y., Lyu, K., Rawat, A.S., Menon, A.K., Rostamizadeh, A., Kumar, S., Kagy, J.F., Agarwal, R.: Distillspec: Improving speculative decoding via knowledge distillation. arXiv:2310.08461 (2023)